Grant-free Non-orthogonal Multiple Access for IoT: A Survey

Muhammad Basit Shahab, *Member, IEEE*, Rana Abbas, *Member, IEEE*, Mahyar Shirvanimoghaddam, *Senior Member, IEEE*, and Sarah J. Johnson, *Member, IEEE*

***Abstract*—Massive machine-type communications (mMTC) is one of the main three focus areas in the 5th generation (5G) of wireless communications technologies to enable connectivity of a massive number of internet of things (IoT) devices with little or no human intervention. In conventional human-type communications (HTC), due to the limited number of available channel resources and orthogonal resource allocation techniques, users get a transmission slot by making scheduling/connection requests. The involved control channel signaling, negligible with respect to the huge transmit data, is not a major issue. However, this may turn into a potential performance bottleneck in mMTC, where huge number of devices transmit short packet data in a sporadic way. To tackle the limited radio resources and massive connectivity challenges, non-orthogonal multiple access (NOMA) has emerged as a promising technology that allows multiple users to simultaneously transmit their data over the same channel resource. This is achieved by employing user-specific signature sequences at the transmitting devices, which are exploited by the receiver for multi-user data detection. Due to its massive connectivity potential, NOMA has also been considered to enable grant-free transmissions especially in mMTC, where devices can transmit their data whenever they need without the scheduling requests. The existing surveys majorly discuss different NOMA schemes, and exploit their potential, in typical grant-based HTC scenarios, where users are connected with the base station, and various system parameters are pre-defined in the scheduling phase. Different from these works, this survey provides a comprehensive review of the recent advances in NOMA from a grant-free connectivity perspective. Various grant-free NOMA schemes are presented, their potential and related practical challenges are highlighted, and possible future directions are thoroughly discussed at the end.**

***Keywords*—Non-orthogonal multiple access (NOMA), massive machine-type communications (mMTC), internet of things (IoT), random access (RA), grant-free transmission.**

I. Introduction

The Internet of Things (IoT) in recent years has emerged as a revolutionary transformation, where almost every physical device is expected to be connected to a communication network through a wired/wireless channel [1]–[3]. Emerging services such as remote monitoring and real-time multi-device control (e.g., connected cars/homes, moving robots, and sensors) represent some prominent examples of the IoT framework. A high proportion of these services demonstrate autonomous communication, where data transmission between various devices, and with the underlying network, takes place with little or no human intervention [4].

A. IoT Traffic Framework

International telecommunication union (ITU) and third generation partnership project (3GPP) have defined three network usage scenarios by considering the variety of connected devices and their diverse quality of service (QoS) requirements. These include enhanced mobile broadband (eMBB), massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC) [5]. The eMBB use case typically refers to the human-type communications (HTC) or human-to-human communications, where the number of devices is less, communication is majorly downlink (DL), and the data size per device is large. On the contrary, mMTC and URLLC use cases of the IoT framework exhibit very different features from the HTC.

In massive IoT or mMTC (e.g., devices reporting to cloud, smart buildings, logistics tracking, and smart agriculture), some key traffic characteristics are: majorly uplink (UL), very small transmit-data size per device, extremely high energy efficiency requirement, partially/fully autonomous communication, and most importantly sporadic transmission [6], [7]. However, these massive IoT use cases may have some degree of tolerance on data reliability and latency constraints. On the contrary, in critical IoT or URLLC (e.g., tele-surgery, intelligent transportation, etc.), highest priority is given to the data reliability and low latency [8]–[10]. For instance, in critical medical applications, end-to-end latency for robot aided and augmented reality assisted surgeries is targeted to be less than 2ms and 750μs respectively [9]. Considering these diverse service requirements, significant upgrades to the existing communication technologies are needed to support these IoT use cases.

The major focus of this work is on mMTC, which is enabled through machine-type communications (MTC), also known as machine-to-machine (M2M) communications, where data transmission between various MTC devices (MTCDs), and with the underlying network, takes place with little or no human intervention [11]–[13]. According to recent Cisco annual internet report (2018-2023), there will be around 29.3

Manuscript received November XX, 2019; revised April XX, 2020; accepted May XX, 2020. Date of publication XXXX; date of current version XXXX. This research was supported by the Australian Research Council under the Discovery Projects grant DP180100606. (*Corresponding author: Muhammad Basit Shahab.*)

Muhammad Basit Shahab and Sarah J. Johnson are with the School of Electrical Engineering and Computing, The University of Newcastle (UoN), Callaghan NSW 2308, Australia (e-mail: basit.shahab@newcastle.edu.au; sarah.johnson@newcastle.edu.au).

Rana Abbas and Mahyar Shirvanimoghaddam are with the School of Electrical and Information Engineering, The University of Sydney, Camperdown NSW 2006, Australia (e-mail: rana.abbas@sydney.edu.au; mahyar.shirvanimoghaddam@sydney.edu.au).

TABLE I: List of abbreviations

3GPP	Third generation partnership project
5G	Fifth generation wireless systems
BOMA	Building block sparse-constellation based orthogonal multiple access
BS	Base station
CDMA	Code division multiple access
CoF	Compute-and-forward
CoSaMP	Compressive sampling matching pursuit
CP	Cyclic prefix
CS	Compressive sensing
CS-MUD	Compressive sensing multi-user detection
CTU	Contention transmission unit
DL	Downlink
eMBB	enhanced mobile broadband
ESE	Elementary signal estimator
FDS	Frequency domain spreading
FEC	Forward error correction
GA	Grant acquisition
GOCA	Group orthogonal coded access
HARQ	Hybrid automatic repeat request
HTC	Human-type communications
IDMA	Interleave division multiple access
IGMA	Interleave-grid multiple access
IoT	Internet of things
JMPA	Joint message passing algorithm
LCR	Low complexity receiver
LCRS	Low code rate spreading
LDS	Low density spreading
LDS-CDMA	Low density spreading code division multiple access
LDS-OFDM	Low density spreading orthogonal frequency-division multiplexing
LDS-SVE	Low density spreading signature vector extension
LPMA	Lattice partition multiple access
LSSA	Low code rate and signature based shared access
LTE	Long term evolution
LTE-A	Long term evolution Advanced
MA	Multiple access
MAC	Multiple access channel
MAP	Maximum a posteriori probability
MCS	Modulation and coding scheme
MIMO	Multi-input multi-output
ML	Machine learning
MPA	Message passing algorithm
MTC	Machine-type communications
mMTC	Massive machine-type communications
MTCD	Machine-type communications device
MUD	Multi-user detection
MUSA	Multi-user shared access
M2M	Machine-to-machine
NCMA	Non-orthogonal coded multiple access
NOCA	Non-orthogonal coded access
NOMA	Non-orthogonal multiple access
NR	New radio
OMA	Orthogonal multiple access
PDMA	Pattern division multiple access
PD-NOMA	Power domain non-orthogonal multiple access
PIC	Parallel interference cancellation
PRACH	Physical random access channel
QoS	Quality of service
RA	Random access
RACH	Random access channel
RAN	Radio access network
RAR	Random access response
RB	Resource block
RDMA	Repetition division multiple access
RIS	Reconfigurable intelligent surface
RSMA	Resource spread multiple access
SAMA	Successive interference cancellation aided multiple access
SCMA	Sparse code multiple access
SDMA	Spatial division multiple access
SINR	Signal-to-interference-plus-noise ratio
UE	User equipment
UL	Uplink
URLLC	Ultra-reliable low-latency communication

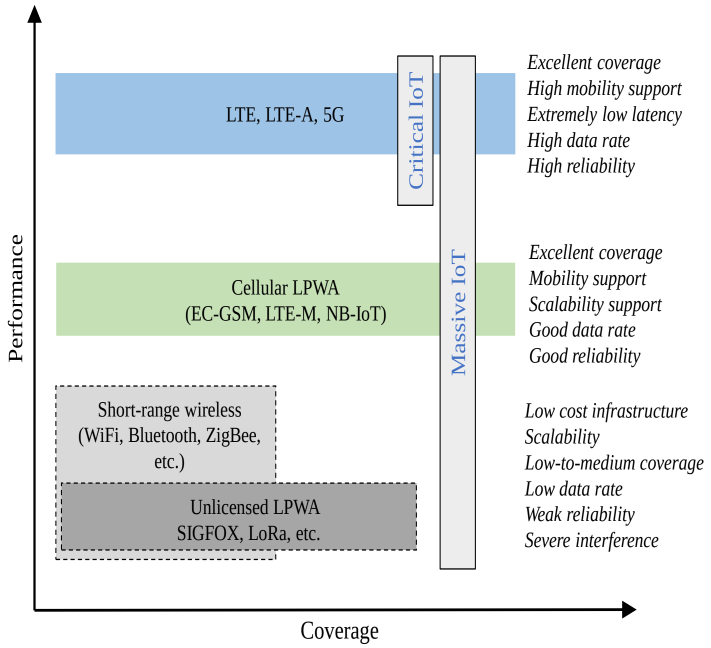


Fig. 1: Wireless technologies for IoT applications; adapted from [8, Fig. 3].

billion networked devices by 2023, where the number of M2M devices is predicted to be 14.7 billion; a share of around 50% of the global connected devices and connections [14]. Similarly, Ericsson predicts that the number of devices connected to communication networks will reach 31.4 billion by 2023, out of which more than 60% will be MTCDs and other IoT connections [15]. Considering these massive mMTC devices, a dramatic shift from the current protocols, mostly designed for HTC, will be needed.

B. *Wireless Connectivity Options*

Fig. 1 depicts the variety of wireless connectivity options to support different use cases in the IoT framework. The basic classification of these technologies for different IoT use cases is carried out by taking network coverage and performance criterion into consideration [8]. It is expected that a large share of these devices will be facilitated by short-range radio technologies, such as WiFi, Bluetooth, and Zigbee, while a significant share will be enabled through wide area networks (WANs) that are majorly facilitated by cellular networks. Connectivity through cellular networks will be provided by the 3GPP technologies, including global system for mobile communications (GSM), wideband code division multiple access (WCDMA), long term evolution (LTE), LTE advanced (LTE-A), and the upcoming 5G. These technologies operate on the licensed spectrum and are primarily designed for high quality mobile voice and data services. An evolution of these technologies is being carried out for low-power IoT applications, where the key challenges are low device cost, long battery life, indoor connectivity and regional coverage, scalability, and diversity.

TABLE II: Summary of existing surveys on NOMA in a chronological order (some prominent schemes covered, grant-free access not discussed in detail): ✓ → covered in detail, ● →briefly introduced, X→ not-covered.

Survey	Grant-based MA signature NOMA schemes				Grant-free NOMA schemes			
	Power domain	Spreading	Scrambling	Interleaving	MA Signature	Compute and forward	Compressed sensing	Machine learning
Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends, Sep. 2015 [16]	✓	✓	X	X	X	X	X	X
A survey: Several technologies of non-orthogonal transmission for 5G, Oct. 2015 [17]	✓	✓	X	X	X	X	X	X
Analysis of non-orthogonal multiple access for 5G, Feb. 2016 [18]	✓	✓	X	X	X	X	X	X
Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges, 2nd quarter 2017 [19]	✓	X	X	X	X	X	X	X
Uplink multiple access schemes for 5G: A survey, Jun. 2017 [20]	✓	✓	✓	✓	X	X	X	X
A survey on non-orthogonal multiple access for 5G networks: Research challenges and future directions, Oct. 2017 [21]	✓	●	X	X	X	X	X	X
Nonorthogonal multiple access for 5G and beyond, Dec. 2017 [22]	✓	●	X	X	X	X	X	X
A survey and taxonomy on nonorthogonal multiple-access schemes for 5G networks, Jan. 2018 [23]	✓	●	X	X	X	X	X	X
Modulation and multiple access for 5G networks, 1st quarter 2018 [24]	✓	✓	X	X	X	X	X	X
Embracing non-orthogonal multiple access in future wireless networks, May 2018 [25]	✓	X	X	X	X	X	X	X
Uplink nonorthogonal multiple access technologies toward 5G: A survey, Jun. 2018 [26]	✓	✓	✓	✓	●	X	X	X
A survey of non-orthogonal multiple access for 5G, 3rd quarter 2018 [27]	✓	✓	✓	✓	●	X	X	X
Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications, Dec. 2018 [28]	X	✓	✓	✓	●	X	X	X
A survey of rate-optimal power domain NOMA schemes for enabling technologies of future wireless networks, Sep. 2019 [29]	✓	X	X	X	X	X	X	X
Interplay between NOMA and other emerging technologies: A survey, Dec. 2019 [30]	✓	X	X	X	X	X	X	X

3GPP has made significant improvements to meet the requirements of emerging massive IoT applications, which lead to a range of cellular low-power wide area solutions. They include a) extended coverage GSM (EC-GSM), which is achieved through new data and control channels mapped over legacy GSM, b) narrow-band IoT (NB-IoT), which is a self-contained carrier with a system bandwidth of 200 kHz and is enabled on an existing LTE network, and c) LTE for MTC (LTE-M), providing new power-saving functionalities to LTE. Overall, some noticeable improvements by 3GPP to enable massive IoT are lower device cost by reducing peak data rate and memory requirements, improved battery life by using power saving mode and discontinuous reception, and better coverage. While significant improvements focused on mMTC have been made, some major challenges related to massive connectivity still need to be tackled, as explained next.

C. Massive Connectivity

The massive connectivity challenge to support mMTC can be broadly split into two categories.

1) Orthogonal/Non-orthogonal Multiple Access: One primary challenge to provide connectivity to these devices is the limited number of available channel resources. The situation is exacerbated by the fact that radio resource allocation in existing multiple access (MA) techniques, i.e., orthogonal MA (OMA), is non-overlapping in nature, such that a radio resource can be allocated to only a single device/user. Considering a large number of devices and limited available radio resources, such OMA based resource allocation becomes a performance bottleneck. In this context, non-orthogonal MA (NOMA) is identified as a promising technology to provide massive connectivity. It works on the non-orthogonality principle, where multiple users can simultaneously transmit their

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

data over the same radio resource block (RB) by employing user-specific MA signature sequences [16]–[30], which can be exploited by the receiver for efficient data recovery, hence creating a window of opportunity to enable massive connectivity over limited radio resources.

In recent years, various NOMA techniques have been proposed by academia and industry. Most of these techniques are analyzed in literature from a HTC perspective, where a limited number of users are considered to be already connected to the base station (BS), and capacity maximization is the key target goal. Moreover, considering centralized scheduling, the users and BS are assumed to know almost everything about each other i.e., number of multiplexed users, MA signature sequences, modulation and coding scheme (MCS), channel state information, etc. Moreover, perfect system synchronization is assumed. These assumptions may not be applicable to mMTC scenarios, and need further investigation.

2) Grant-based/Grant-free Access: Although NOMA can provide massive connectivity by loading multiple users over a RB, another challenge is the way each device accesses a channel resource. In existing wireless networks, each device requests a data transmission slot via a contention-based random access (RA) process, which (in LTE/LTE-A) is identified as a major performance bottleneck, and a source of excessive delay and signaling overhead. Considering sporadic mMTC traffic, a stepwise shift towards grant-free/contention-based communication is inevitable, where devices can transmit data as per their needs without going through the RA/grant process, or by merging RA and data transmission. In this context, grant-free/contention-based transmission using NOMA is considered as a promising solution.

D. Contributions and Organization

Table. II provides a summary of how some prominent NOMA schemes (discussed later in Sec. III) have been covered in recent surveys [16]–[30]. While these works extensively cover grant-based NOMA, detailed information from a grant-free perspective is still lacking. In this context, this survey primarily focuses on the grant-free access mechanisms coupled with NOMA techniques, which so far have only been briefly touched upon in the existing surveys [26]–[28]. The problem for grant-free NOMA is fundamentally different than grant-based NOMA schemes as the set of active users is unknown to the receiver. Many challenges arise with grant-free NOMA including, but not limited to, design of contention transmission units that are chosen at random by users, synchronization with minimal signaling overhead, blind detection of users and channel estimation, collision detection and resolution of data, etc. The major contributions of this article are highlighted as follows, whereas the organization is shown in Fig. 2.

- To the best of our knowledge, this is the first comprehensive review of grant-free NOMA schemes. The article categorizes these schemes into four main types, namely, (1) MA signature sequence-based (2) compute-and-forward-based, (3) compressive sensing-based, and (4) machine learning-based. Moreover, besides a thorough review of the recent advances in these schemes, their potential and possible future directions are also discussed in detail.

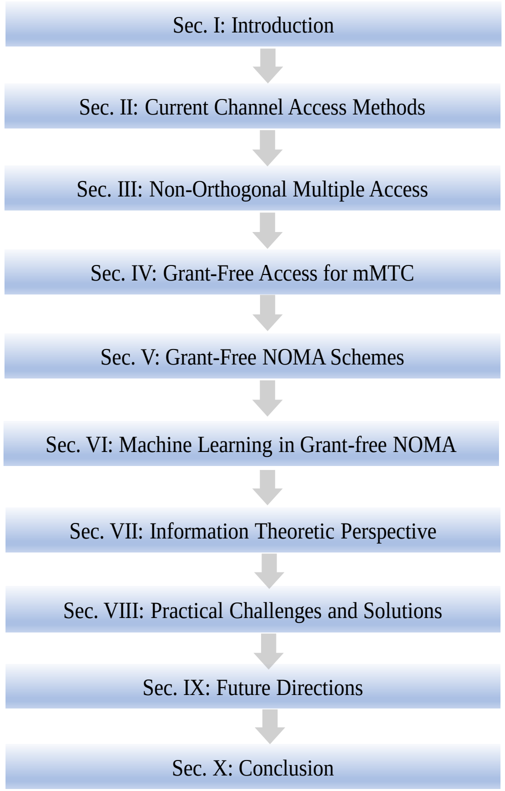

Sec. I: Introduction
Sec. II: Current Channel Access Methods
Sec. III: Non-Orthogonal Multiple Access
Sec. IV: Grant-Free Access for mMTC
Sec. V: Grant-Free NOMA Schemes
Sec. VI: Machine Learning in Grant-free NOMA
Sec. VII: Information Theoretic Perspective
Sec. VIII: Practical Challenges and Solutions
Sec. IX: Future Directions
Sec. X: Conclusion

Fig. 2: Organization of the paper.

- A comprehensive review of the recent efforts in information theory on RA is then provided, followed by discussion on the non-negligible gap between the existing bounds and the practical schemes existing to date. Information theoretic perspective of grant-free access through NOMA is then discussed.
- Finally, a comprehensive list of challenges for the use of grant-free NOMA for mMTC/IoT is presented, and some possible future directions are discussed.

II. Current Channel Access Methods

3GPP has recently been working on the design of network architecture, services, and optimization measures for mMTC. This is being done by considering a huge number of mMTC devices with a variety of service requirements such as ultra-low energy consumption, low cost, and the coexistence of both mMTC and HTC applications.

A. Channel Access Methods in LTE/LTE-A

In existing wireless networks, radio resources (e.g., time, frequency) are allocated orthogonally to connected devices. Therefore, in LTE/LTE-A, the entity requesting access to the cellular network has to first go through a contention-based process over the physical random access channel (PRACH) to get aligned with the BS, which (in LTE/LTE-A) has been identified as a major performance bottleneck, and a source of excessive delay and signaling overhead [31], [32]. The process was originally intended to be used for multiple purposes i.e., initial access of a user not connected to the BS,

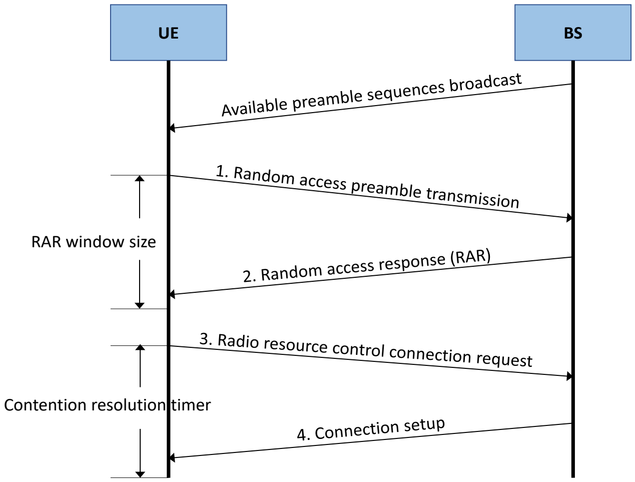

UE
BS
Available preamble sequences broadcast
1. Random access preamble transmission
RAR window size
2. Random access response (RAR)
3. Radio resource control connection request
Contention resolution timer
4. Connection setup

Fig. 3: RACH of LTE/LTE-A networks.

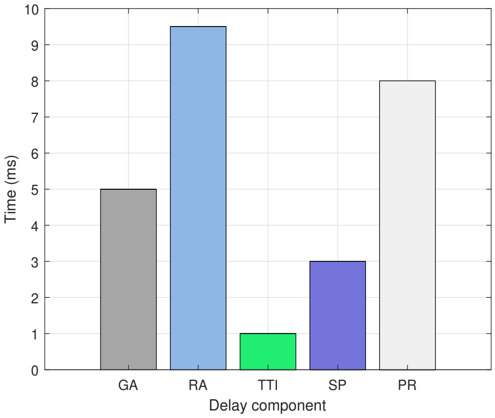


Fig. 4: Delay sources in LTE (Rel. 8); GA means grant acquisition of a connected user to transmit data, RA is RACH plus GA for a new user, TTI is the transmit time interval, SP is signal processing, and PT is packet re-transmissions.

UL synchronization of connected users, data transmission or acknowledgment response, handover management, etc., [33]. The LTE standard prescribes that PRACH access be performed using a four-way handshake, which is contention-based (the users can initiate the access process whenever they want).

To enable PRACH access, devices are initially informed about available PRACH resources through a broadcast from BS. This is followed by the four-step RA channel (RACH) handshake procedure. Any device which is already aligned with the BS can make a grant acquisition (GA) request to transmit its data when it needs. The RACH of LTE/LTE-A [34] is shown in Fig. 3, and the steps are summarized below.

1) **Preamble transmission:** Every device randomly chooses one out of the available preambles, and sends it to the BS, which estimates the transmission time of each device from its detected preamble.
2) **Random access response:** For each detected preamble, BS sends a RA response (RAR) message with information about the radio resource allocated to the device and timing advance information for synchronization. If a device does not receive RAR within a predefined waiting time (RAR waiting window size) or gets a RAR with no information about its request, it postpones the access attempt to the next RACH opportunity.
3) **Radio resource control connection request:** Each device that gets a successful RAR from the BS makes a radio resource control (RRC) connection request by sending its temporary terminal identity to the BS through the Physical UL shared channel (PUSCH).
4) **RRC connection setup:** BS sends information of allocating resources to all devices that have gained access by specifying their terminal identity.

B. Related Performance Issues

The RA process in Fig. 3 comes with many problems, especially latency. In Fig. 4, some sources of delay in the LTE Release 8 are summarized. Furthermore, there can be other delays due to the core network, such as queuing delay due to congestion, packet re-transmission delay caused by upper layers, etc. It can be seen from Fig. 4 that the RA process causes a latency of around 9.5 ms, which is too high for certain IoT use cases mainly with strict latency requirements [32].

The RA process is feasible for a small number of devices. However, if PRACH is congested due to a large number of MTC and/or HTC devices attempting access simultaneously, the signal-to-interference-plus-noise ratio (SINR) observed at the receiver (BS) may be reduced to the extent that messages cannot be detected, and consequently, many of the access attempts fail; this is denoted as the outage condition [35]. Excessive preamble collisions and re-transmissions result in problems like network congestion, unexpected delays, packet loss, radio resource wastage and high energy consumption. Excessive overhead is another major issue as significant resources are spent only to establish a connection for communicating very small sized mMTC traffic data [36]. For instance, to transmit 100 B of data, around 59 B of overhead in UL and 136 B of overhead in DL would be required for signal transmissions [37]. In LTE-A, the collision rate of the RA requests is around 10 percent and the signaling overhead is about 30-50 percent of the payload size [38].

C. Proposed Access Method Modifications for mMTC Traffic

In the past several years, a number of techniques to improve the performance of mMTC traffic by alleviating outage due to congestion have been suggested. For instance, grouping the traffic into different classes for which RA process may be temporarily delayed or blocked, as done in access class barring technique [39]. Differentiation among traffic classes can be achieved by allowing different back-off windows for different classes of users [40]. Similarly, predefined access attempt slots can be allocated based on the user grouping [41], or suitable polling scheme can be applied [42]. Moreover, PRACH resources (such as preamble sequences and RA slots) can be separated and dynamically allocated [43].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

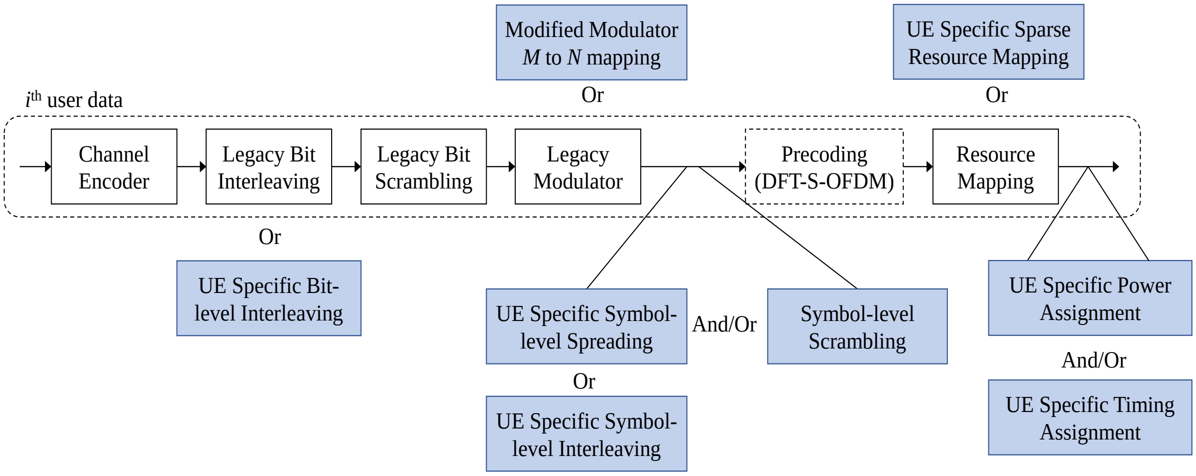


Fig. 5: General structure of NOMA transmitter processing.

In addition to finding ways for reducing outage due to congestion, optimizing the RACH procedure has also been investigated in recent years by 3GPP radio access network (RAN) working group 1 (WG1) for the new radio (NR). For instance, it has been agreed that NR should support multiple RACH preamble formats with shorter/longer preamble lengths. Five different PRACH formats for the RA preamble parameters with different lengths are currently available in LTE [44]. Details about preamble design, and corresponding RAR improvements, for NR PRACH can be found in [45]–[48].

Currently, there is only one RACH timeline applied to all scenarios and use cases in LTE. However, as NR supports various services and use cases with different latency requirements, one RACH timeline might not be efficient. Hence, the need for RACH timeline design with variable lengths for a variety of use cases is critical, and is addressed in [49]. In addition to these, 2-step RACH procedure is also proposed for consideration in 3GPP NR [50]. The use cases for such a RACH procedure are users with typically intermittent small packet transmissions, access in unlicensed spectrum, cases where UL timing advance is not needed, etc., [51].

While improvements in grant-based access procedures are ongoing, these might not be enough to cater the mMTC traffic. A primary reason behind this is the number of devices and their transmission pattern in mMTC, which is completely different from the HTC, for which the RA approaches were basically designed. mMTC involves a huge number of devices with sporadic transmissions, which may congest the PRACH if many devices simultaneously try to get access. Accordingly, grant-based approaches will fail under such conditions. To tackle this issue, grant-free access has gained significant research interest in recent years, where devices can transmit their data whenever they want in an ”arrive and go” manner. Moreover, to facilitate grant-free access, the use of NOMA has been jointly agreed by both academia and industry.

III. Non-orthogonal Multiple Access

To enable massive connectivity and to deal with the limitations of existing MA schemes, need for the design of new wireless technologies is inevitable. In this context, NOMA has emerged as a potential MA technique for 5G and beyond. The core idea of NOMA is to superimpose the data streams of multiple users over the same RBs by employing user-specific signature patterns that facilitate multi-user detection (MUD) at the BS (see references in Tab. II for details). Hence, NOMA provides massive connectivity and high spectral efficiency compared to existing OMA schemes [52]–[56]. It has been agreed that NOMA schemes should be investigated for diverse 5G usage scenarios [57], and 5G should target to support UL NOMA at least for mMTC [58]. In this context, a brief insight into the MA signature design for NOMA is presented next.

A. *Signature Design for NOMA*

The overloading performance and complexity of the MUD receiver at the BS depends on the design of MA signatures in NOMA [28]. To achieve this, various operations, such as linear spreading, interleaving, scrambling, and multi-dimensional modulation can be employed at the transmitter as depicted in Fig. 5. The blocks in black and white reuse the current NR design, while new blocks with specification impact are highlighted in blue. These operations can be applied at the bit-level and/or symbol level for efficient MA signature design. Moreover, integration of multiple operations can also be used to develop multi-layer MA signatures.

B. *Prominent NOMA Schemes*

Based on the different MA signatures, various NOMA schemes have been proposed by academia and industry in recent years, and are briefly summarized in Table. III. These schemes have been categorized in existing literature on the basis of MA signatures such as spreading, scrambling, interleaving, and power domain [59], [60], as shown in Table. III. Some of these schemes have also been considered as candidate solutions for supporting UL of mMTC [59], [60]. Interested readers can look into the references provided in Table. III and the surveys listed earlier in Table. II for details.

TABLE III: Various NOMA schemes proposed by academia and industry

	NOMA Schemes		Description	Receiver Type
1	Power based	PD-NOMA [61]–[84]	● Users transmit their data with different power levels over the same RB. ● BS exploits the received power difference to perform MUD using SIC receiver.	SIC
2	Spreading based	LDS-CDMA [85]–[87]	● Inspired by CDMA, where users share a RB through unique user-specific spreading sequences. ● Uses LDS or sparse spreading sequences to limit interference on each chip unlike classic CDMA.	MPA
3		LDS-OFDM [88]–[90]	● An integration of LDS-CDMA and conventional OFDM. ● Symbols are first multiplied with LDS sequences, and mapped onto different OFDM subcarriers. ● More fit for wideband than LDS-CDMA, and achieves significant performance improvement.	MPA
4		SCMA [91]–[98]	● Developed from basic LDS-CDMA. But, bit-to-constellation mapping and spreading are combined. ● Each user has its own unique codebook for bit to codeword mapping. ● Codebooks are built using multi-dimensional constellation mapping, that provides constellation shaping gain and more diversity.	MPA
5		SAMA [99]	● Different from other LDS schemes, spreading sequences in SAMA have variable sparsity. ● Hence, data of users are spread over different number of resources, providing more diversity.	MPA
6		PDMA [100]–[104]	● Similar to SAMA, sparsity of spreading sequences employed by users is variable. ● Moreover, users are multiplexed in multiple domains i.e., power, space, code, or their combination.	MPA/SIC
7		LDS-SVE [105]	● Based on original LDS, the idea is to design larger user-specific signature (spreading) vectors. ● For example, concatenating two element signature vectors of a user into a larger signature vector. ● Such signature-vector-extension can further exploit diversity gain.	MPA
8		MUSA [106]	● Dense-spreading scheme. Uses complex spreading codes with short length and SIC receiver. ● Increased pool of spreading codes due to the real and imaginary parts.	SIC
9		NCMA [107], [108]	● Spreading codes are obtained through Grassmannian line packaging problem.	PIC
10		NOCA [109]	● Based on LTE defined low correlation sequences as spreading codes.	SIC
11		FDS [110]	● Directly spreads the modulation symbols with multiple orthogonal or quasi-orthogonal codes.	SIC
12		LCRS [110]	● Direct spreading of modulation symbols with multiple orthogonal codes.	SIC
13		GOCA [111]	● The spreading sequences in GOCA have two-stage structure; dual spreading sequences. ● Non-orthogonal sequences are used to enable group separation. ● Moreover, device separation within a group through a set of orthogonal sequences..	SIC
14	Srambling Based	RSMA [112], [113]	● Uses combination of very low rate FEC and long scrambling sequences with good correlation. ● Moreover, different interleavers can also be optionally used.	SIC
15		LSSA [114]	● Uses low rate FEC or moderate one with repetition and bit/symbol-level permutation patterns.	SIC
16	Interleaving Based	IDMA [115], [116]	● Unique user-specific bit-level interleaving is used; may include low rate FEC and/or repitition.	ESE
17		IGMA [117]	● Uses bit-level interleavers and symbol-level grid mapping patterns for user separation.	ESE
18		RDMA [111]	● Employs symbol-level cyclic-shift repetition patterns to design device-specific signatures. ● Contrary to IDMA, random interleaver is not used in RDMA.	SIC
19	Others	SDMA [118]	● Uses unique user-specific channel impulse responses to multiplex users. ● Due to variety of channel impulse responses, large number of users can be supported. ● However, accurate channel impulse response estimation is needed at the base station.	PIC
20		LPMA [119]	● Uses multi-level lattice superposition codes to allocate different code levels to multiplex users. ● With multiple degrees of freedom in multiplexing, LPMA provides more flexible/diversity.	SIC
21		BOMA [120]	● User multiplexing by attaching information from good channel user to symbols of bad user. ● The bad channel gain user applies coarse constellation with large minimum distance.	LCR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

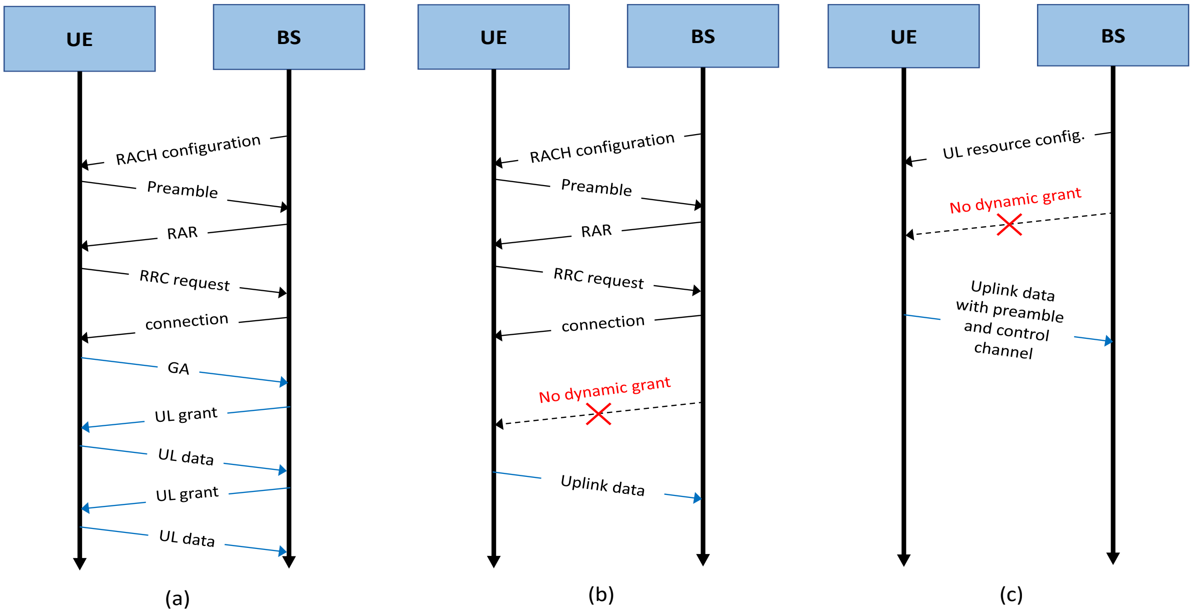


Fig. 6: UL access as described in [121] (a) RACH-based grant-based, (b) RACH-based grant-free, (c) RACH-less grant-free.

IV. Grant-Free Access for mMTC

As mentioned in Sec. III, many variants of NOMA have been proposed by academia and industry in recent years. However, analysis of these schemes in existing literature mainly considers scheduling/grant-based scenarios, where spreading sequences, interleaving patterns, and/or transmission powers of different users are predefined by the BS. However, the major drawback of this is the excessive signaling overhead, which makes grant-free NOMA inevitable.

A stepwise proposal towards grant-free communication using NOMA-based user multiplexing for 3GPP NR was presented in [121], with the following four steps.

1) RACH-based grant-based OMA as the starting point or baseline UL transmission scheme. To deal with the problems of RA process discussed in Sec. II-B, and motivated by several improved RA strategies discussed in Sec. II-C, new RA methods can be designed to minimize collision and overload problems.
2) RACH-based grant-based NOMA schemes. In [122], a novel RA strategy to enable multiple MTCDs to transmit over same RB is developed. The presence of preambles is detected using timing advance information. Moreover, through power control, BS can detect the number of devices which have selected the same preamble. This enables BS to perform RACH signaling for a group of devices instead of each individual, which significantly reduces the signaling overhead in mMTC. These devices can then be allowed to communicate using grant-based NOMA transmission over the same data channel.
3) RACH-based (synchronous UL) grant-free NOMA to reduce the UL grant overhead. The MTCDs can perform the RACH, but once they get aligned with the BS, further transmissions of data need no GA, and are grant-free NOMA based.
4) RACH-less (asynchronous UL) grant-free NOMA. In this case, the MTCDs transmit their data without carrying out any RA or GA process.

Fig. 6 shows an illustration of RACH-based grant-based, RACH-based grant-free, and RACH-less grant-free transmissions for UL of 3GPP NR [121]. There are two possible options of resource allocation in grant-free transmissions. First option is that MTCD's radio resource is pre-configured by BS or pre-determined, while the second option is that MTCD performs random resource selection itself [123]. In both cases, when the MTCD wants to transmit data, it makes no further grant requests, and is termed as a grant-free transmission. Fig. 7 illustrates a grant-free/contention-based arrive and go transmissions using OMA, where some i^{th} and j^{th} users transmit whenever they have data packets [124]. In case both users transmit in the same grant-free slot, a collision is said to have taken place due to OMA, and they re-transmit their data later using random back-off procedure.

The state graph of a grant-free transmission is shown in Fig. 8. If user has no data to transmit in its buffer, it stays in a sleep state; otherwise, it would wake up, synchronize using reference signals, and acquire some necessary system information. Before direct grant-free transmission, preamble may be transmitted for UL synchronization to facilitate detection at receiver. Furthermore, some MA information could be implicitly indicated by the preamble, such as spreading signature, locations of radio resources, and the timing of re-transmissions. With this information, collisions can be detected, and the blind detection complexity of BS can also be greatly reduced. In general, grant-free UL transmission schemes need to ensure that transmission parameters, identification of the MTCD for purpose of decoding data (e.g., knowledge of spreading code, interleaver, etc.), synchronization, and channel estimation can be determined or detected by the BS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

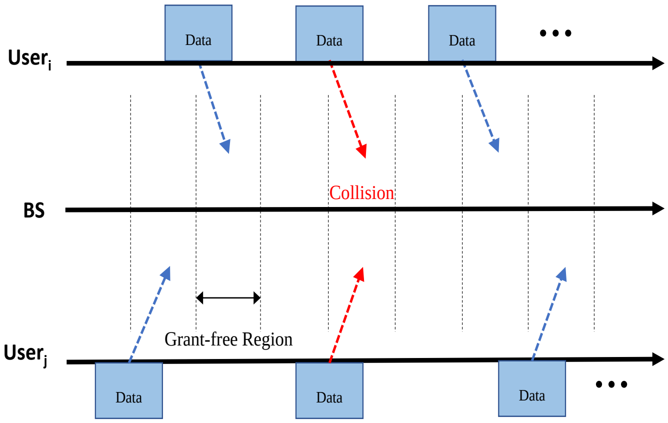


Fig. 7: Grant-free contention based OMA transmission.

Grant-free transmission using OMA schemes will cause a tremendous amount of collisions due to limited available resources as depicted in Fig. 7, where two users transmitting in the same slot collide and need re-transmissions. However, in case of NOMA, use of different signature patterns will avoid these collisions as BS can distinguish between multiple users transmitting over the same RB through their signature patterns. Hence, the basic features of grant-free/contention-based UL NOMA are a) transmission from a MTCD does not need the dynamic and explicit scheduling grant from BS, and b) multiple MTCDs can share the same time-frequency resources through NOMA. It was agreed that grant-free NOMA schemes, where MTCDs can send their data without going into any explicit dynamic grant process, are well-suited for mMTC [125]. The devices can transmit their data whenever they want, which can reduce signaling overhead and end-to-end latency [126]. Hence, it was decided that 3GPP NR should aim to support UL grant-free/contention-based transmissions at least for mMTC [127], [128].

As mentioned earlier, depending on whether RACH is present or not, we have RACH-based/RACH-less grant-free UL NOMA [128]. Collectively, they come under the umbrella of grant-free transmission. In RACH-based grant-free NOMA, once all users perform the RACH, grant-free transmission can occur in a more synchronized manner. In RACH-less grant-free NOMA, to reduce the signaling overhead, RACH can be completely eliminated, and the data transmission phase starts whenever a user has packets to transmit. The problem of grant-free NOMA is fundamentally different than grant-based with many challenges including, but not limited to, design of contention units that are chosen at random by users for transmission, synchronization with minimal signaling, and most importantly blind data detection. Here, the set of active/transmitting users is unknown to the BS. Therefore, it has to blindly perform active user detection, channel estimation and data recovery. These tasks can either be performed in different phases or combined together as a one-shot joint operation. Moreover, the receiver also needs to identify and resolve any collisions due to users transmitting data over the same channel resource with same MA signature.

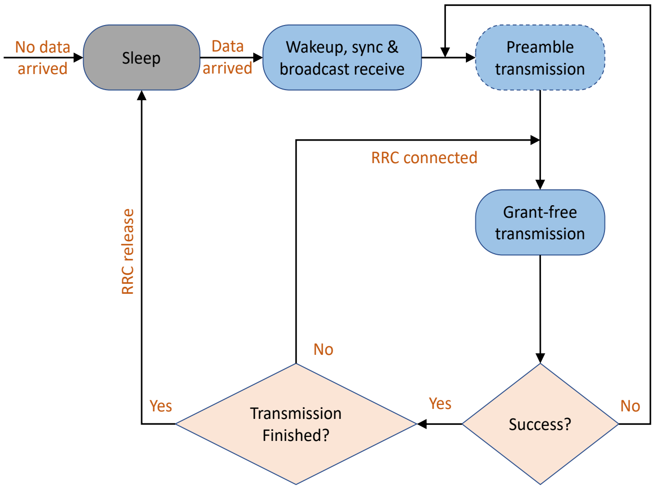


Fig. 8: Illustration of grant-free UL transmissions; reproduced from [126, Fig. 1].

V. Grant-Free NOMA Schemes

We categorize grant-free NOMA schemes into four main classes: (1) MA signature based, (2) compute-and-forward (CoF) based, (3) compressive sensing (CS) based, and (4) machine learning (ML) based. The first three classes are thoroughly reviewed in this section, whereas ML based grant-free NOMA is presented exclusively in the next section. A summary of these schemes is provided in Table. IV.

A. MA Signature-based Grant-free NOMA

In grant-free UL NOMA, BS may not have complete information about multiplexed users, with various other transmission parameters also unknown/partially-known. To enable grant-free access, a contention-based MA resource is defined, which comprises of a physical resource (a time-frequency block) and a MA signature, which may include at least one of the following; codebook/codeword, sequence, interleaver and/or mapping pattern, demodulation reference signal, power-dimension, spatial-dimension, preamble, etc. [127]. For MA signature selection, one option is that MTCD performs random selection and the other option is that MTCD's signature is pre-configured/pre-determined. Through these MA signatures, various grant-free NOMA transmissions can be enabled. Some prominent MA signature (i.e., spreading, scrambling, interleaving, etc., discussed in Sec. III) based grant-free UL strategies and MUDs are discussed in this section.

1) Power-based Grant-free NOMA: Power domain NOMA (PD-NOMA) works on the principle of superimposing multiple users' data streams over the same RB using power levels. When multiple users transmit their message signals using different powers, the BS exploits the received power difference to perform MUD using a SIC receiver. PD-NOMA has received significant research interest in recent years. The scheme has recently been included in ATSC 3.0, a forthcoming digital TV standard, under the name layered-division-multiplexing. Moreover, it is considered for 3GPP LTE enhancements under

TABLE IV: Grant-free UL NOMA schemes

Grant-free NOMA schemes	Description
MA signature based	• A contention-based MA resource is defined consisting of a time-frequency block and a MA signature. • The MA signature can be scrambling/spreading/interleaving based, and therefore includes at least one of the following; codebook/codeword, sequence, interleaver and/or mapping patterns, power dimension, sptaial dimension, preamble, pilot, etc. • Multiple users can transmit in a grant-free manner by using any MA resource, where MUD at BS is achieved by exploiting the MA signatures.
CoF based	• Users encode their messages with two concatenated channel codes; one for error correction, and one for user detection. • The first, inner code, is to enable the receiver to decode the sum of all codewords. • The second, outer code, is to enable the receiver to recover the individual messages of users that participated in the sum.
CS based	• In UL grant-free mMTC, the inherent sparsity of user activities could be used to solve the MUD problem by using CS. • Exploiting the low user activity ratio, CS techniques enable the BS to handle more users. • As blind MUD at BS needs to jointly perform user activity and data detection in grant-free UL, this activity detection can be achieved through CS-MUD by exploiting the sporadic nature of mMTC transmissions. • Moreover, user-specific signature patterns enable the MUD to distinguish between these active users at the receiver.
ML based	• Inspired by powerful capabilities of ML to look for patterns in data for making best possible, nearly optimal, decisions. • BS is trained to extract the features of the NOMA signal received as a result of multiple active/transmitting users.

the name multi-user superposition transmission. However, these applications, and indeed most of the existing work on PD-NOMA is grant-based.

As the scheme heavily depends on the received power difference among users, the effectiveness of such a scheme may be limited for grant-free solutions in the absence of closed-loop power control. Therefore, maintaining sufficient power difference among signals of multiple users/MTCDs received at the BS is a huge practical challenge. PD-NOMA generally needs a deterministic near-far situation for successful operation. But in grant-free access, the random near-far problem would make the SIC receiver for PD-NOMA difficult to implement [67], [68], [125]. A detailed analysis of the effects of channel gains and power allocations of multiplexed users on their error rate and capacity for a PD-NOMA scenario is performed in [69]. It has been shown that the channel gains of users and their power allocation is critical for the performance of SIC receiver. However, some PD-NOMA based UL grant-free schemes do exist in literature such as integration of ALOHA or slotted-ALOHA protocols with PD-NOMA [129]–[131]. In these schemes, the BS adaptively learns about the number of active devices by using multi-hypothesis testing, and a novel procedure enables the transmitters to independently select distinct power levels.

2) Spreading-based Grant-free NOMA: The core idea behind spreading-based NOMA schemes is the use of user/device-specific low cross-correlation real/complex-valued spreading sequences to enable non-orthogonal transmission and MUD. The schemes are further classified into low density spreading (LDS) and non-LDS. While both sequence types can be considered, the former is preferable because of its ability to efficiently mitigate multi-user interference. Various spreading based schemes were summarized earlier in Tab. III.

In general, all MA signature-based NOMA schemes can be customized to achieve grant-free UL transmission. We consider spreading-based NOMA schemes in this part. The spreading code can be designed with either long or short sequence. Users randomly select the sequence from a predefined codebook set or a spreading sequence resource pool. To enable grant-free access, a contention transmission unit (CTU) is defined. The CTU is a basic building block of a predefined region within the time-frequency plane for grant-free/contention-based transmissions, and may consist of several fields including radio resources, reference signals, and spreading sequences [132]. A CTU differs from others in any fields, and these differences can be exploited by receiver for efficient MUD. A MTCD, which has data to transmit, randomly selects a CTU and transmits its data packet accordingly. For mMTC, different MTCDs may choose the same radio resource, but different fields of the CTU still facilitate the BS in efficient MUD.

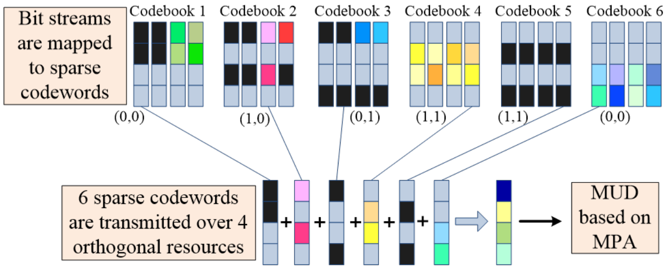


Fig. 9: Resource mapping of SCMA: 6 users, 4 subcarriers; from [133, Fig. 3].

To elaborate further, let us consider conventional SCMA as an example. Developed using LDS-CDMA, the original bit stream of each user is directly mapped to a codeword chosen from its own dedicated codebook. SCMA codewords are sparse, i.e. only few of their entries are non-zero. The key difference between LDS-CDMA and SCMA is that SCMA relies on multidimensional constellations for generating its codebooks. All SCMA codewords have a unique location of non-zero entries. An illustration of resource mapping of SCMA is shown in Fig. 9, by considering 6 users, 4 resources,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

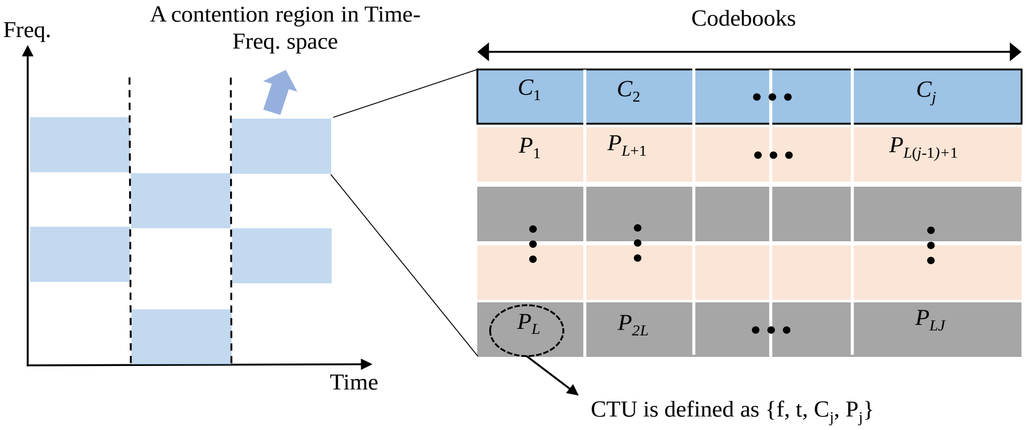


Fig. 10: An illustration of CTU in a time-frequency space; adapted from [132, Figs. 2 and 3].

and sparsity of 2 (each user transmits data over 2 out of 4 resources). The maximum number of codebooks J that can be generated is a function of K (non-zero entries in a codeword) and N (codeword length). Selection of K non-zero positions within N elements is simply a combination problem. The maximum number of such combinations is given by the binomial coefficient $J = \binom{N}{K}$. Hence, for the example shown in Fig. 9 with a 4-dimensional complex codebook ($N = 4$) and 2 non-zero entries ($K = 2$), a set of $J = \binom{4}{2} = 6$ codebooks is generated, where each user selects one codeword from its codebook. Each user maps its two bits (b_1, b_2) to a codeword. The data is then spread over the subcarriers. In this case, the data streams of multiple users are overlaid with codewords from different codebooks. As there are 6 possible codebooks, 6 users can be multiplexed over 4 subcarriers (150 percent loading).

In order to provide massive connectivity, a contention-based/grant-free SCMA scheme is proposed in [132]. In this context, a CTU is designed, as shown in Fig. 10. This CTU is a combination of time, frequency, SCMA codebook, and pilot sequence. There are J unique codebooks defined over the time-frequency resources (RBs). For each codebook, there are L associated pilot sequences, making it a total of $L \times J$ unique combinations. In this way, there is a resource pool of $L \times J$ CTUs in the given time-frequency region. Multiple users/MTCDs may reuse the same codebook, and transmit at the same time. As codebooks go through different wireless channels, the MPA based detector can still detect these MTCDs data carried over the same codebook, as long as different pilot sequences are used. Moreover, the receiver can estimate channels of different MTCDs with different pilots. Therefore, the number of active users at a specific time slot can be potentially more than J. In this way, codebook reuse can help to increase the effective overloading factor and the number of connections to enable mMTC.

The user-to-CTU mapping rule can be defined, so that each MTCD can select the CTU itself. For instance, a MTCD can choose the CTU index as CTU$_{\text{index}}$=MTCD$_{\text{ID}}$ mod N_{CTU}; the CTU index that a MTCD chooses to transmits its data over is a function of the MTCD ID and the resource pool of CTUs N_{CTU} [132]. In case different MTCDs choose the same CTU for transmission, a collision will occur, which can be resolved through random back-off procedure.

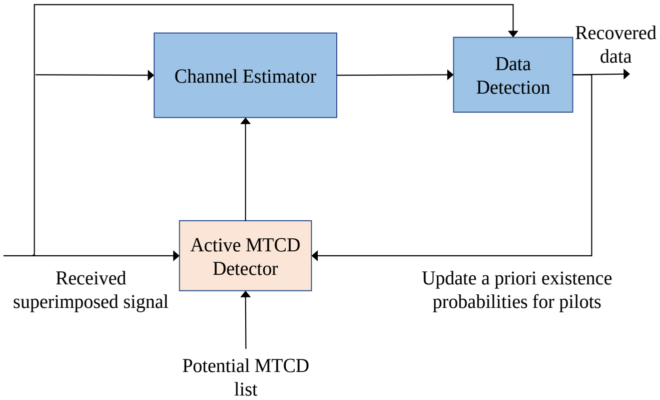


Fig. 11: Illustration of MUD for grant-free NOMA.

It is important to note that the whole time-frequency region does not need to be contention supportive. For practical purposes, only a portion of UL bandwidth is configured as contention regions, while the other portion can be used for regular scheduled UL data transmissions. The coexistence of contention-based NOMA and scheduled access is supported by various studies in academia and industry. This is based on the fact, that in addition to the MTCDs, there are a comparatively smaller number of devices using eMBB services, where scheduled access has been shown to be quite efficient [134]. The size and number of access regions are therefore dependent on many factors e.g., expected number of MTCDs and/or applications, etc. An illustration of CTU regions in time-frequency space is shown in Fig. 10.

As aforementioned, in conventional SCMA, number of codebooks J depends on spreading factor and non-zero entries, i.e., $J = \binom{N}{K}$. Hence, by varying the spreading factor and degree of sparsity, significant increase in the codebooks pool

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

is possible. For instance, for $N = 8$ and $K = 4$, 70 codebooks can be generated. Correspondingly, the number of CTUs in the contention region increases manifold, thereby improving the performance of grant-free access.

In order to perform MUD, different types of receivers are proposed in literature. In this context, performance comparison of SCMA with three types of receivers i.e., MPA, MPA with SIC, and MPA with Turbo decoder, is shown in [135]. Moreover, by considering the use of CTUs and massive number of MTCDs with sporadic grant-free transmissions, a blind detection-based receiver is proposed in [136], as shown in Fig. 11. The receiver basically consists of two components, where the first one identifies active MTCDs to narrow down the list of potentially active MTCDs, and second one i.e., data detection block referred also as joint data and active codebook MPA (JMPA) detector to decode the active MTCDs with no knowledge of active codebooks. The first component (active MTCDs detector) acts as a pre-filter to narrow down the list of potential active MTCDs to control the complexity and efficiency of reception. This can be achieved through some efficient CS-based techniques detailed in the next section. Based on a short list created by active MTCD detector through pilot symbols, the channel estimator now needs to estimate only the channels of these identified active MTCDs. In addition to receiver, efficient design of codebooks for SCMA may also facilitate in enhancing the performance of grant-free transmission with efficient detection. In this context, a detailed analysis of variable overloading and robustness to codebook collision for SCMA is provided in [137].

Similar to the case of SCMA, spreading sequences in the CTUs can reuse the sequences designed for other spreading based schemes such as LDS-SVE, PDMA, MUSA, etc., to facilitate grant-free transmissions. As it was mentioned earlier, a hard collision may occur if multiple MTCDs select exactly the same CTU. In these situations, these MTCDs may be distinguished and detected only if they have distinctive channel gains. However, from this perspective, an important solution is to enlarge the pool of spreading sequences, as done in MUSA, which is one of the candidate techniques for UL grant-free transmission [138]. Multiple random non-orthogonal complex spreading codes with short length constitute a pool in MUSA, from which each user/MTCD can randomly choose one. It is to be noted that for the same user, different spreading sequences may also be used for different symbols in order to improve the performance via interference averaging. All spreading symbols of MTCDs are transmitted over the same time-frequency resources. At the receiver, codeword level SIC is used to separate data from different users. Complex spreading sequences maintain lower cross-correlation than traditional pseudo random noise-based sequences due to the additional freedom of the imaginary part. It should be noted that spreading sequences of MUSA are different from LDS and do not have low density property.

As long spreading sequences used in traditional CDMA have relatively low cross-correlation, these codes combined with SIC for grant-free communication would cause processing complexity, delay and the error propagation in the receiver. Hence, short spread codes with relatively low cross-correlation are suitable for grant-free UL MUSA [133]. In this context, the family of complex spreading codes is a suitable option, as they are short due to the design freedom with real and imaginary parts. These spreading sequences are specifically designed to cope with heavy overloading of users, and to facilitate simple SIC on the receiver side. Moreover, in order to enable grant-free transmission and minimize the overhead of control signaling, users choose their spreading sequences locally/autonomously, without coordination by BS [139]. A RACH-free grant-free MUSA transmission model is shown in Fig. 12. Whenever a MTCD has data to transmit, it randomly chooses a spreading code, and transmits data without any RACH, GA, or power control. At the BS, blind channel estimation and MUD using SIC receiver is performed.

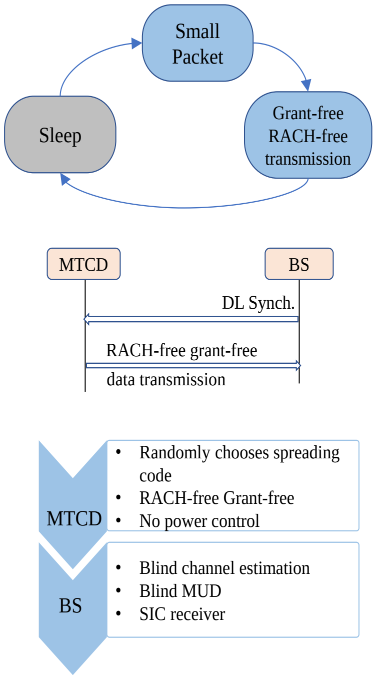


Fig. 12: RACH-free grant-free transmission based on MUSA.

Overall, some practical challenges like collisions and receiver computational complexity are well addressed in MUSA. As the elements of spreading sequence are not binary, codewords do not need to be sparse, and more elements can be used by the spreading code in MUSA. In this way, a large pool of spreading codes can be generated, reducing the collision probability compared to SCMA and PDMA. Meanwhile, the complexity of blind MUD increases significantly as the pool size grows. Therefore, the pool size should be set to reasonable values in order to achieve massive connections while limiting the complexity of blind MUD [139]. Hence, the design of spreading sequence is crucial to MUSA since it determines the interference between different users and system performance. Details about MUSA, its spreading process, and grant-free transmissions are provided in [133], [140], [141].

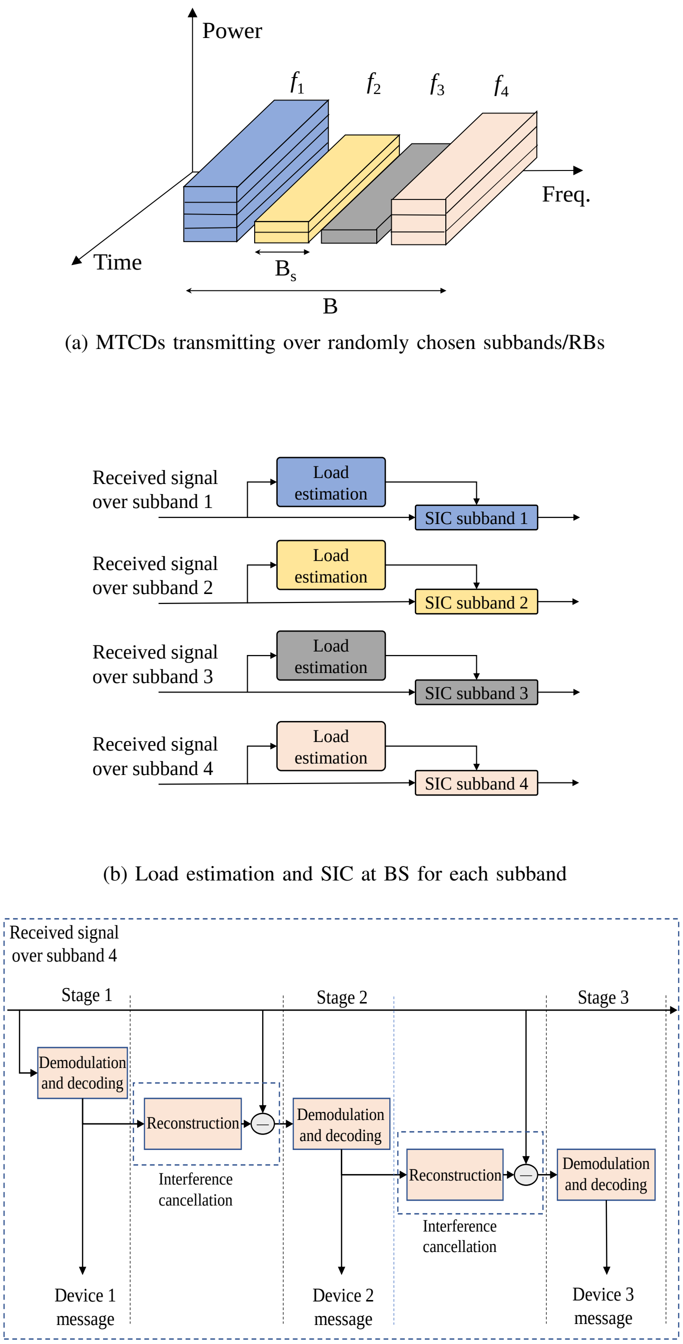
While most of these works consider an ideal SIC-based MUD to show the performance bounds of various spreading codes, a realistic receiver for grant-free MUSA is discussed in [142], [143]. As mentioned earlier, the important assumptions for ideal receiver, such as the active users' spreading codes and their fading channel may not be easily known at the BS before decoding. Fortunately, taking full advantage of the characteristics of grant-free accessing, e.g. inherent random near-far phenomenon and MUSA's special features, e.g. short spreading code, blind MUD with very high performance is proposed and analyzed in [142], [143]. The blind MUD investigated for MUSA consists of the following components. A SIC receiver is used to take full advantage of the near-far phenomenon usually observed in grant-free systems. Blind estimation is suggested by making full use of the characteristics of the spreading codes and the received signal. With blind estimation for the user with the highest post-SINR in current SIC stage, the decoding performance of the detected user can be guaranteed. Moreover, blind advanced channel estimation using pilot/reference signal is also suggested, where long preambles are not needed in the channel estimation to save the overhead. With the help of the advanced channel estimation and SIC, the post-SINR of the next detected user signal is enhanced and can be successfully decoded. It was shown through simulation results that the block error rate of MUSA, with blind SIC receiver, is smaller than 0.1 even with 300 percent user loading.

3) Scrambling/interleaving-based Grant-free NOMA: Scrambling based NOMA schemes use a combination of very low-rate forward error correction (FEC) codes and user/device-specific scrambling sequences to enable non-orthogonal transmissions and MUD at the BS. Similarly, interleaving based NOMA schemes use different user/device-specific interleavers along with repetition and/or low-rate FEC codes for device separation. The key characteristic of interleaving based NOMA schemes is the use of different interleavers to distinguish between multiplexed users. Further details for various scrambling/interleaving based NOMA schemes are provided earlier in Tab. III.

In order to enable grant-free transmissions in scrambling/interleaving based NOMA schemes, a similar procedure as that described for spreading-based schemes can be employed. While scrambling/interleaving based schemes are also considered for use to provide grant-free access in UL of mMTC [133], there is very limited work in the literature to exploit their potential. In this context, there is still a need for research to analyze the overloading capability and error rate performance of scrambling/interleaving based NOMA schemes in grant-free scenarios.

4) Other Grant-free Signature-based NOMA Schemes: Some other grant-free transmission schemes motivated by signature-based NOMA are also proposed in academia. For instance, in [144], a Raptor code-based NOMA for mMTC is proposed with random data packet arrivals. In the considered system, MTCDs do not need to perform RA to get a transmission slot or network access. Instead, the RA and data transmission phases are combined over randomly selected subbands to minimize overhead. The BS, being unaware of the number of MTCDs multiplexed over a subband performs load estimation and SIC for MUD.

The steps and details of Random NOMA strategy are summarized as follows.


Power
f_1
f_2
f_3
f_4
Freq.
Time
B_s
B
Received signal over subband 1
Load estimation
SIC subband 1
Received signal over subband 2
Load estimation
SIC subband 2
Received signal over subband 3
Load estimation
SIC subband 3
Received signal over subband 4
Load estimation
SIC subband 4
Received signal over subband 4
Stage 1
Stage 2
Stage 3
Demodulation and decoding
Reconstruction
Interference cancellation
Demodulation and decoding
Reconstruction
Interference cancellation
Demodulation and decoding
Device 1 message
Device 2 message
Device 3 message

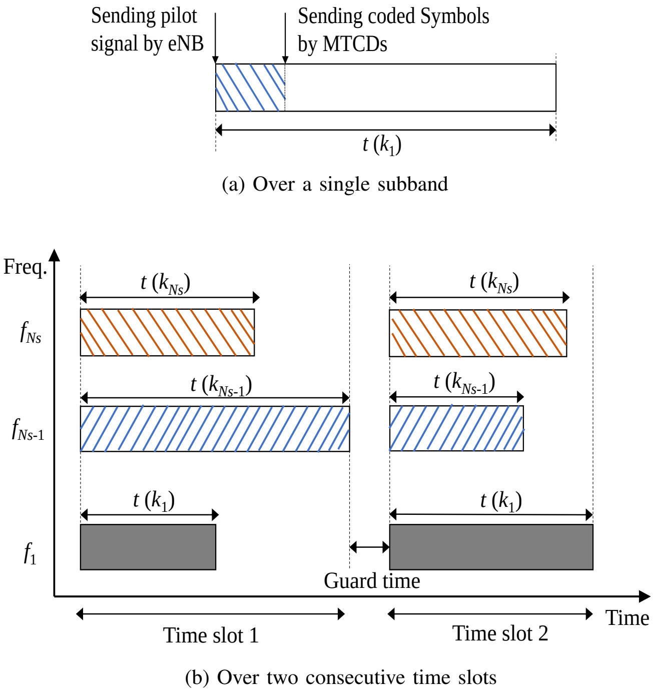
(a) MTCDs transmitting over randomly chosen subbands/RBs

(b) Load estimation and SIC at BS for each subband

(c) SIC process for users over a subband

Fig. 13: Illustration of grant-free Random NOMA for mMTC.

1) Initially, BS broadcasts a pilot signal over each subband at the beginning of a time slot.
2) Each MTCD with data for transmission randomly chooses a subband from a set of available subbands, and listens to the pilot signal transmitted by BS over that subband. Correspondingly, the MTCD estimates channel over that subband. Moreover, the MTCD also randomly chooses a seed for its random number generator from a set of M_s available seeds.
3) Each active MTCD, after randomly selecting a subband and seed, attaches its unique terminal ID to its message,



(a) Over a single subband

(b) Over two consecutive time slots

Fig. 14: Time slot duration of grant-free Random NOMA.

and encodes using a Raptor code constructed from the selected seed, and transmit the codeword over the selected subband.

4) When BS receives superimposed message signals of various users over a subband, it performs load estimation followed by SIC for MUD. The SIC order is such that BS starts decoding with the first seed and removes interference of it, followed by second seed and so on.

An example scenario is shown in Fig. 13a, where bandwidth is divided into 4 subbands, and users randomly choose a subband for data transmission. The data retrieval at BS is shown in Figs. 13b and 13c, where BS performs load estimation (number of multiplexed users) for each subband, followed by SIC process to separate and recover the data of each user over that subband. All MTCDs are assumed to perform power control such that their received power at the BS is same. In this way, BS can effectively estimate the number of MTCDs multiplexed over each subband by calculating the total received power, as it would be proportional to the number of overlapped MTCDs. Further details about load estimation algorithms can be found in [122].

Due to random subband selection by MTCDs, the achievable rate over each subband is not fixed and depends on the number of randomly overlapped MTCDs. Therefore, the number of coded symbols to be transmitted over each subband is also random. Fig. 14 shows the variable length of each subband in two consecutive time slots. It is important to note that each time slot duration will be mainly determined by the subband with highest number of active MTCDs, as its maximum achievable rate would be lower than the rest. Therefore, a fixed-rate code cannot be used in all time instances. Accordingly, use of Raptor codes is proposed in [144].

Raptor codes are rateless and can generate as many coded symbols as needed by BS. They have a random structure represented by a bipartite graph, which depends upon a pseudo random generator's seed. By using the same seed, BS can reproduce the same bipartite graph as the MTCD and decode the message. In case multiple MTCDs select the same seed while transmitting over the same subband, their transmitted code structure will be exactly the same. In this case, a collision is said to have taken place, as the BS cannot differentiate between these users due to same structure of received codewords. However, the probability of such collision is still far less than the conventional RA collision. The work in [144] compared the average number of successful supported MTCDs by the random NOMA scheme with access class barring. It was shown that the proposed scheme supported significantly larger number of devices compared to access class barring.

Improvisations in random NOMA may completely avoid collisions. For instance, assume that each device always uses a unique user-specific seed determined in advance, then the collisions can be completely eliminated. However, this may still be complicated when number of MTCDs is incredibly large; need for massive seed pool and complex detection process at BS by trying such massive seeds for data recovery of multiplexed users over each subband. The performance limits of random NOMA scheme for massive cellular IoT are comprehensively discussed in [145]. In [146], a novel framework for grant-free NOMA is developed, where collisions between any number of users over a radio resource does not entail for all simultaneously transmitting multiplexed users, and is treated as interference to the remaining received signals. Moreover, as PD-NOMA is less discussed from a grant-free UL perspective, a novel multi-level grant-free scheme is proposed in [147]. The layers correspond to different codebooks and different power levels. The users first choose a layer randomly, which corresponds to a codebook. Users in each layer utilize the same codebook and perform power control such that they have the same received power level at the BS. To achieve this, users choose between a predetermined set of power levels. At the receiver, the sum of codewords over each layer are first separated from remaining layers, symbol by symbol. Then, the codewords of the same layer are jointly decoded at the BS.

B. CoF-based Grant-free NOMA

NOMA schemes explained in the previous section are based on the MA signature design by employing user-specific sequences i.e., spreading, scrambling, interleaving, or any combination of these, to enable MUD at the receiver. The general concept of this MA signature design was earlier illustrated in Fig. 5. Besides the traditional MA signatures, another approach towards grant-free NOMA is proposed in [148], [149], where the principle of CoF [150] is employed. The scheme relies on codes with a linear structure, specifically nested lattice codes. The linearity of the codebook ensures that integer combinations of codewords are themselves codewords. A destination is free to determine which linear equation to recover [148]. The concept of network coding in CoF can be interpreted as a conversion of a network into a set of reliable linear equations. Inspired by this, in CoF based grant-free NOMA, users encode their messages with two different channel codes where one is used for error correction and the other is used for user detection. The latter is chosen such that the sum of K or less distinct codewords is unique. Correspondingly, at the receiver side, the BS has to first decode the sum of the received codewords. In situations when the BS cannot recover the sum correctly, all the transmitted data is lost, which is similar to the existing MA signature-based grant-free NOMA schemes discussed earlier.

A channel use is divided into multiple sub-blocks, where each active user randomly chooses a sub-block, over which it transmits. All users encode their messages using the same codebook C, which are then modulated. The codebook C is developed as a concatenation of two codes. The first one is an inner binary linear code, whose purpose is to enable the receiver to decode the modulo-2 sum of all codewords transmitted within the same sub-block, which can be referred as the CoF phase. The second code is an outer code, whose purpose is to enable the receiver to recover the individual messages of users that participated in the modulo-2 sum. This recovering of the individual messages from their modulo-2 sum can be referred as the binary adder channel (BAC) phase. The success probability of the CoF phase in [149] is independent of the actual number of users that transmitted within the same sub-block. However, the outer code in [149] is designed such that if at most M users use the channel over a particular sub-block, it is possible to determine the individual messages from their modulo-2 sum, essentially with zero error probability.

The design of an inner code to be used in the CoF phase reduces to that of finding codes that perform well over a binary input memoryless output-symmetric channel, for which many off-the-shelf codes can be used. For the outer codes used in the second phase/stage, they can be constructed from the columns of T-error correcting Bose-Chaudhuri-Hocquenghen (BCH) codes [151]. As mentioned, at the receiver side, the BS has to first decode the sum of the received codewords. When the BS cannot recover the sum correctly, all the transmitted data is lost which is similar to the existing works on signature-based grant-free NOMA schemes. Moreover, in CoF based grant-free NOMA [148]–[150], a closed loop power control is used with the extension to open-loop power control being not straight-forward. A similar approach with similar challenges is considered in [152]–[154] using physical layer network coding where users transmit to a multi-antenna BS.

C. CS-based Grant-free NOMA Schemes

The MUD problem in grant-free NOMA transmissions was briefly discussed previously in Sec. V-A. Due to the sporadic nature of data transmission by devices in mMTC, active device detection and data recovery are of prime importance, where the BS needs to identify some K out of T total users that were active and transmitted data in a particular time slot. To this end, the use of CS algorithms for receiver design has gained significant research interest in grant-free scenarios. It is important to note that the term CS-based grant-free NOMA does not imply that CS is also a NOMA scheme. NOMA transmissions here are still enabled through any of the MA signatures described previously, whereas CS is only used for efficient MUD. As there is significant work on CS-based receiver designs for grant-free NOMA, we have summarized all the related works in this separate section.

CS is an efficient signal processing technique that exploits the sparsity of a signal to recover it from far fewer samples than required by the Nyquist criteria. In current wireless systems, the number of active users is usually much smaller than the total number of available users in the system, even during busy hours [155]. This characteristic also applies to mMTC scenarios. Thus, in UL grant-free NOMA systems, the inherent sparsity of user activities could be used to solve the MUD problem by using CS algorithms. Exploiting the low user activity ratio, CS techniques enable the BS to handle more users [156], because CS makes it possible to recover the desired signals from far fewer measurements than the total signal dimensions if signal sparsity is assumed.

NOMA with CS-MUD is considered as a promising candidate to enable grant-free UL NOMA for mMTC. As blind MUD at the BS needs to jointly perform user activity and data detection in grant-free UL, this activity detection can be achieved through CS-MUD by exploiting the sporadic nature of mMTC transmissions, thereby allowing the CS-based algorithms to detect user activity. Moreover, user-specific signature patterns enable the MUD to distinguish these active users at the receiver. CS-MUD based NOMA schemes allow the users/MTCDs to transmit their data whenever they want without any control signaling.

Considering a fewer number of active users compared to the actual number, two models are used in literature to formulate NOMA as a CS problem. The first model is termed as a single-measurement vector based CS (SMV-CS), where a one shot transmission is considered by taking the received signals as a vector $\mathbf{y}$, which consists of superimposed signals of active users. The received vector $\mathbf{y}$ is product of a vector $\mathbf{d}$ consisting of data symbols of the users, and a sensing matrix $\mathbf{A}$, which contains the influences of channel and spreading matrices. In SMV-CS, when the number of users increases, the size of this sensing matrix $\mathbf{A}$ becomes huge, which leads to poor sampling matrix properties and hence limits the scalability of the system in terms of detection speed. To deal with this, multiple-measurement vector based CS (MMV-CS) considers the received signal as a matrix $\mathbf{Y}$, which is a product of two matrices, i.e., a sensing matrix $\mathbf{A}$ and a data symbols matrix $\mathbf{D}$, where each row of $\mathbf{D}$ contains data frame of a single user and each column represents the symbols from all users at a time instant. This is done to reduce the size of the sensing matrix $\mathbf{A}$. Hence, compared with the SMV-CS model, MMV-CS can better mitigate higher complexity due to the growing number of users. These models have been actively used to represent the sparse spreading-based NOMA in UL mMTC scenarios [157]–[159]. Moreover, in order to perform CS-MUD, various algorithms exist in literature. These algorithms can be categorized as maximum *a posteriori* probability (MAP) based algorithms and greedy algorithms [160], and are used frequently for CS-MUD.

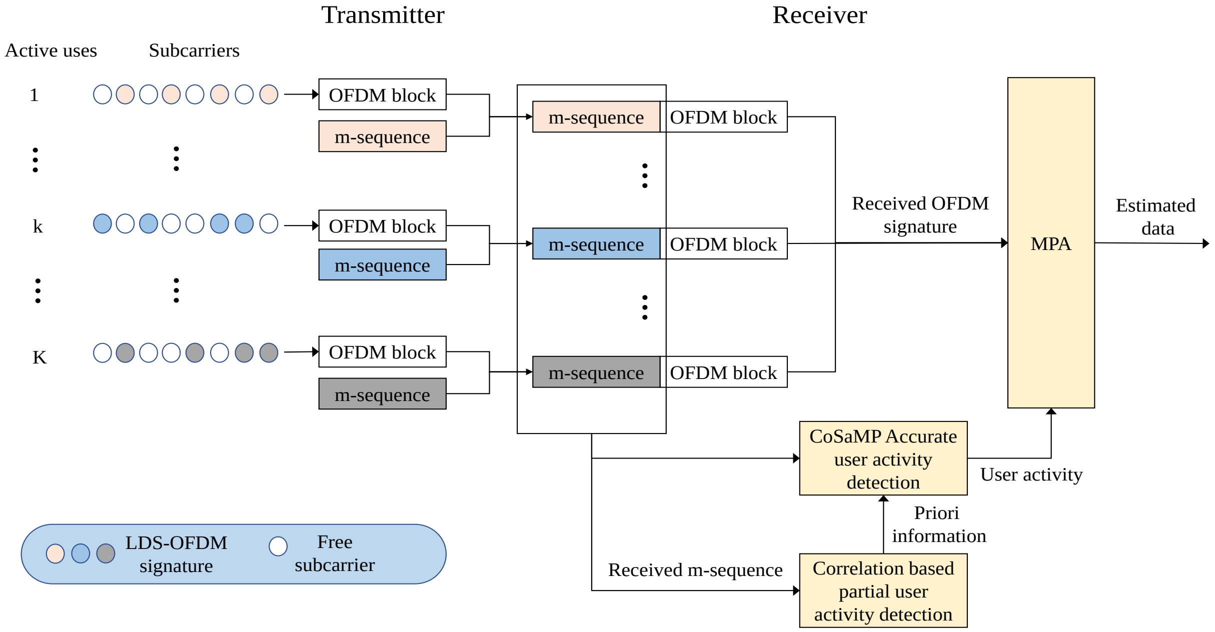


Fig. 15: CS-based time-frequency joint NOMA transceiver; reproduced from [164, Fig. 3].

In the UL grant-free NOMA systems, the current near-optimal MUD based on MPA was discussed earlier to approximate the optimal MAP detection. The receiver assumes that the user activity information is exactly known, which is impractical, yet challenging, due to any of the massive users randomly entering or leaving the system. In this context, a joint use of CS and MPA for designing a CS-MPA detector to realize both user activity and data detection for LDS-based UL grant-free NOMA is proposed in [161]. The sparse signal recovery algorithms in CS can be used to realize user activity detection by identifying the positions of non-zero elements, for which a compressive sampling matching pursuit (CoSaMP, [162]) algorithm is used due to its low complexity and excellent robustness to noise. CoSaMP is one of the many CS algorithms, and is a common method applied to detect nonzero elements in a sparse signal by using the intrinsic sparsity of signals. After the activity detection through CoSaMP, by making full use of the sparsity of LDS structure, low complexity MPA based receiver is employed. Similarly, by observing a structured sparsity of user activity in mMTC, a low-complexity MUD based on structured CS for NOMA to further improve the signal detection performance is proposed in [163].

In [164], the CS-MPA of [161] is optimized by incorporating a two-stage CS-based activity detection for LDS-OFDM UL grant-free NOMA as shown in Fig. 15. In the first stage, a correlation-based activity detection is carried out. The approximated support obtained at the first stage is then fed into the second stage which executes the CoSaMP algorithm. In addition to these MUD receivers, by considering the variations in the sparsity level of the multi-user signals, a switching mechanism between the CS-MUD and the classical MUD can be adopted as proposed in [165]. Similarly, as the sparsity of active users varies from time to time, a low complexity dynamic CS-based MUD is proposed in [166]. This is based on the idea that although users can randomly access/leave the system, some users generally transmit their information in adjacent time slots with a high probability, which leads to the temporal correlation of active user sets in several continuous time slots [167]. By exploiting this temporal correlation, the estimated active user set in a particular time slot is used as the initial set to estimate the transmitted signal in the next time slot in the dynamic CS-based MUD.

In addition to CS, the user activity detection can also be realized through other algorithms and schemes. For instance, three different algorithms are presented in [136] for active pilot detection namely, channel estimation-based algorithm, focal underdetermined system solver (FOCUSS), and expectation maximization (EM). Furthermore, a sparsity-inspired sphere decoding (SI-SD) based blind detection algorithm for grant-free UL SCMA is proposed in [168]. By introducing one additional all-zero codeword, each user's status and data can be jointly detected, thus avoiding the redundant pilot overhead, and achieving the MAP detection. Similarly, an improved detection-based group orthogonal matching pursuit (DGOMP) MUD is proposed in [169] to facilitate massive grant-free UL SCMA transmissions and reception. Moreover, in [170], a comprehensive study is provided where synchronization, channel estimation, user detection, and data decoding are performed in one-shot.

D. Potential of Grant-free NOMA Schemes and Future Directions

As discussed above, a variety of NOMA schemes exist in literature, and many of them have been considered for use in grant-free scenarios. In what follows, we summarize the potential of these schemes, highlight existing gaps, and suggest some possible future directions.

1) MA Signature-based Grant-free NOMA: PD-NOMA, which in particular has gained significant interest in grant-based DL scenarios, suffers from the lack of power control in the grant-free case. Though there exist some initial works in this context, significant efforts to solve the near-far problem are needed to establish a strong case for the technique to be used in grant-free transmissions. The remaining MA signature-based NOMA schemes, that involve spreading, scrambling, or interleaving of the users' data, are all currently being considered for use in the UL of mMTC. Out of these, spreading-based schemes have been the major focal point of research for grant-free NOMA in recent years. The variety of sequences for data spreading, high overloading capability, and better error rate performance make them one of the most suitable candidates for enabling grant-free access.

Despite the fact that there exists a variety of research work covering different MA signature-based schemes for grant-free access, the work is yet to reach maturity. There are not any widely agreed benchmarks for comparison nor does there exist any generalized set of system parameters over which the schemes can be compared. This limits the comparison between schemes as results in different works remain true for only a certain set of parameters with various assumptions. To tackle this, a unified system parameter benchmark needs to be defined by considering the common properties of various mMTC traffic use cases. Based on these unified parameter settings, different schemes need to be compared for various key performance metrics such as overloading capability, error rate performance, support for asynchronous transmissions, etc. A detailed feasibility study of NOMA for NR in [171] develops a very structured comparison model for various NOMA schemes for different use cases of IoT, thereby providing a good reference for any future works in this area.

Another important area of future research is the integration of different MA signatures for enabling NOMA. The majority of existing works on grant-based/grant-free NOMA focus on analyzing systems by employing a particular MA signature such as power domain or spreading sequence. However, there exists some evidence of using multi-layer schemes, where two MA signatures are simultaneously used by devices to enable NOMA transmissions [147]. The integration of multiple MA signatures is expected to increase the resource pool, thereby reducing the signature collision, and improving the user overloading. However, the work in this area is very limited and needs to be further explored.

2) CoF-based Grant-free NOMA: CoF-based schemes provide another MA signature to be utilized in enabling NOMA transmissions. However, the research work on CoF-based grant-free NOMA schemes, that exists so far in literature, is in its very early stage. Although the schemes do show some benefit for use in grant-free access, a detailed study in this context is required to understand the true potential of CoF-based NOMA schemes.

3) CS-based Grant-free NOMA: It is evident from literature that CS-based NOMA schemes are a very promising candidate for enabling grant-free UL mMTC. However, there still exists some work to be done in this area. For instance, numerous CS-based joint user activity and data detection schemes have been proposed in recent years. The majority of these works are based on the standard sparsity model, which only considers that the multi-user vector is sparse due to the sporadic transmissions. Unfortunately, these works do not impose any underlying structural information (e.g., temporal correlation) to improve the detection efficiency. In practice, we often come across a problem of detecting a sequence of sparse signals which is correlated across time. This is because many devices transmit data in bursts. Hence, these standard solutions only recover each sparse vector independently, which do not take full advantage of all the intrinsic information of UL grant-free NOMA systems. Considering such structural properties can be very beneficial. In this context, the question to answer would be how to jointly reconstruct these sparse signals with temporal correlation to improve the detection efficiency. There exist some recent works that consider consecutive time slot transmissions-based grant-free scenarios, and employ burst-sparsity and frame-wise joint sparsity models to improve the MUD performance. However, the research can be enhanced by exploiting other realistic structural properties of mMTC traffic and their categorization based on the service type.

VI. Machine Learning in Grant-free NOMA

Recent research continues to confirm the powerful capabilities of ML technologies [172] in enhancing the efficiency of transmitter/receiver designs in wireless communications. ML can solve NP-hard optimization problems in a faster, more accurate and robust manner than traditional approaches. Instead of relying on models and equations, ML algorithms look for patterns in the data to make the best possible, nearly optimal, decisions. The robustness of ML algorithms is especially desirable in wireless communications due to the dynamic nature of the networks, whether it is the fast-changing channel states, the dynamic network traffic, or even the network topology and scheduling. For NOMA systems, ML can be applied to several of its NP-hard problems that include (1) attaining channel state information, (2) resource allocation, (3) power allocation, (4) clustering, (5) complex joint decoding, and (6) the fundamental trade-offs among them. This is especially useful in mMTC, as the complexity of these processes grows exponentially with the number of users.

Before we proceed to survey the literature, it is of value to briefly explain some important terminology:

- *Supervised/Unsupervised learning:* Supervised learning aims to learn the mapping function between input data and its respective output (called its label) by minimizing the function approximation error. On the other hand, unsupervised learning aims to extract the inner features of unlabeled data. One example of the latter would be the autoencoder.
- *Reinforcement learning:* Reinforcement learning is based on a reward system where the learner is trained through multiple trial and error attempts to maximize rewards. Video games are a popular example of this paradigm.
- *Deep learning:* Deep learning is a subset of ML that is capable of unsupervised learning from data that is unstructured or unlabeled. It is also known in the literature as deep neural networks.

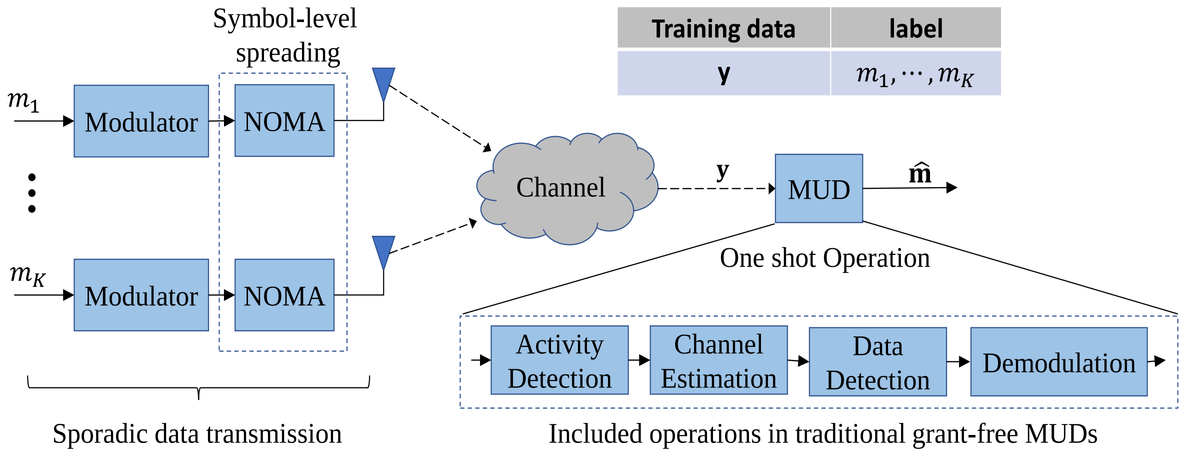


Fig. 16: General principle of machine/deep learning aided MUD for grant-free NOMA.

- *Online/Offline learning:* Offline learning is often also referred to as batch learning. In batch learning, the parameters are updated after consuming a whole batch of data. On the other hand, in online learning, the parameters are updated after each training data learnt.

Although the concept of leveraging ML algorithms to solve communications-related problems can be dated back to almost 20 years ago, it seemed to have little, and possibly no, impact on the way communications was designed and implemented until now. We suspect that the main reason is that information theory, statistics, and signal processing offered very accurate and often flexible models, that allowed for designs with reliable, analytically proven, performance guarantees. Nonetheless, with the development and wide spread of specialized software libraries as well as cheap processing chips, training and testing machine learning models has become much more attractive.

For the physical layer, learning algorithms have been applied to demodulation [173], channel codes [174], CS [175], coherent [176] and blind [177] detection, etc. For higher layers, learning algorithms have been applied to traffic load prediction, control, and channel access schemes. Learning algorithms change the way we fundamentally solve communications problems. Traditionally, to find optimal solutions, our transmitters and receivers are divided into several independent blocks e.g., coding, modulation etc. Similarly, the overall communications system is divided into layers that perform tasks independently e.g., routing, scheduling, resource allocation etc. Such a division, though sub-optimal, allows for the separate analysis and optimization of tasks, and eventually to stabilize systems. However, the approach of learning technologies disrupts this entire process by attempting to accomplish more than one task at once. For example, authors in [178] went as far as demonstrating that it is possible to build a point-to-point communications system whose entire physical layer processing is carried out by neural networks. The work in [179] is another example of this, where unmanned aerial vehicles deployment design and trajectory, amongst other aspects, were optimized jointly through ML.

Here, we review the few, yet growing number of, ML-based papers that can be found in the literature on NOMA thus far. We cover both grant-based as well as grant-free NOMA. In this context, a general model of machine learning based MUD for grant-free NOMA is shown in Fig. 16.

A. ML-based Grant-based NOMA:

In [180], authors use deep learning to solve the high computational complexity of traditional channel estimation and detection algorithms that is mainly due to the fast-changing wireless channel. Long short-term memory (LSTM), a branch of recurrent neural networks, is used to perform automatic encoding, decoding, and channel detection in a DL NOMA system by learning the channel characteristics between the BS and each randomly deployed user. In [181], deep learning is applied to SCMA for codebook design, which is notoriously difficult due to its high dimension. The devised deep neural network can construct codebooks that minimize the block error rate and can adapt to the available set of resources.

In [182], authors propose an online learning detection algorithm for UL NOMA with SIC, where devices are clustered. The goal is to design a partially linear beamforming filter, unlike traditional non-linear beamformers, that is more robust to variations in the environment and dynamic nature of networks with sporadically arriving users. In [183], deep recurrent neural networks learning is used to obtain energy efficient resource allocation for heterogeneous cognitive radio networks. Deep learning has also been applied to power allocation of caching-based NOMA [184]. In [185], deep learning was used for the joint DL resource allocation problem for a multi-carrier NOMA system. Lastly, an unsupervised ML approach is proposed for NOMA in [186] to solve the clustering problem.

B. ML-based Grant-free NOMA:

The work in [187] aims at solving the MUD problem of UL grant-free NOMA through CS without relying on any a priori knowledge, namely the user sparsity level or noise level. The latter two are usually required for the correct termination of the recovery process. Instead, authors adopt statistical and

ML mechanism cross validation (CV) to determine the user sparsity level, mathematically referred to as the model order, and eventually to decide when recovery needs to be terminated. Results show that the proposed algorithm avoids overfitting and underfitting well.

In [188], deep learning is used to solve a variational optimization problem for grant-free NOMA. The neural network model covers encoding, user activity, signature sequence generation, and decoding. Simulation results show that the process can be of very low latency to suit tactile IoT applications. The authors then extend their work in [189] to design a generalized/unified framework for NOMA using deep multi-task learning. In [190], authors propose two MUD schemes, namely, random sparsity learning MUD and structured sparsity learning for synchronous and asynchronous transmissions, respectively. Authors show that even when users do not use pilot signals, the proposed algorithms demonstrate low error rates. Thus, the proposed algorithms can significantly reduce overhead in grant-free access scenarios.

Some other works in [191], [192] focus on user activity detection and channel estimation in grant-free UL NOMA. The work in [191] presents a deep learning-based active user detection scheme for grant-free NOMA. By feeding training data into the designed deep neural network, the proposed active user detection scheme learns the nonlinear mapping between received NOMA signal and indices of active devices. Correspondingly, the trained network can handle the whole active user detection process, and achieve accurate detection of the active users. It is numerically demonstrated that the proposed detection scheme outperforms the conventional approaches by a large margin in both the detection success probability and computational complexity. Similarly, [192] presents a joint user activity and channel estimation method by using block sparse Bayesian learning.

C. Challenges and Future Directions in ML-based Grant-free NOMA

In general, the research work presented thus far in the literature demonstrates very promising performance enhancements to existing systems in terms of faster processing and near-optimal solutions. However, the research work is still too independent to give a clear picture on how things compare, especially when comparing different learning solutions to the same problem. Below, we put forth some issues and concerns related to this line of work, some of which are re-iterated from other authors.

- *Poor choice of benchmarks:* Benchmarks taken from the traditional line of work are often too basic and very old. Researchers need to be comparing against recent cutting-edge solutions proposed in the literature.
- *Poor choice of system models:* If research in this area focuses on simple yet practical system models, we have a chance at comparing our results to the theoretical limits of the system to better understand just how "near" we are to the optimal scenario. This is a crucial point, as it is very difficult to assess what is the best we can get with ML. In other words, performance will almost never be analytically verified which leads us to believe that research on finding theoretical fundamental limits of systems will forever continue to evolve and should not be undervalued.
- *Practical channel statistics:* Systems with unknown channel models need to be better addressed by these approaches [178] in order for learning algorithms to retain their reputation of ability to tune parameters on the fly. So far, all algorithms rely heavily on channel models.
- *User detection solutions need to incorporate collisions (for grant-free NOMA):* All work focuses on user detection and correct estimation of parameters. However, it thus far assumes that users have unique signature sequences and thus collisions are not an issue. However, in massive user settings, assigning unique signature sequences is not practical and collisions are the bottleneck of performance. Thus, collision detection and resolution need to be better addressed in grant-free settings.

VII. Information Theoretic Perspective of Grant-Free NOMA

The capacity bounds of the classical multi access channel (MAC) with a fixed number of users and an infinitely large block length is well understood. However, for many years, information theorists have been hinting at deriving capacity bounds for grant-free channels. For instance, over 30 years ago, [193] sought a coding technology that is applicable for a large set of transmitters of which a small, but variable, subset simultaneously uses the channel. Just under 10 years ago, [194] concluded that when the total number of senders is very large, so that there is a lot of interference, we can still send a total amount of information that is arbitrary large even though the rate per individual sender goes to 0 [194, pp. 546, 547]. With the fast evolution of mMTC, information theorists have amplified their efforts in deriving insightful and easy ways to evaluate capacity bounds that suit these new and pertinent scenarios. The main challenges faced for grant-free NOMA are with the finite block length regime, the random activity of users, and the growth of the number of users with the block length. Below, we provide the main information theoretical studies to date related to grant-free NOMA. Lastly, we shed light on how these works tie in with the design of protocols and codebooks to achieve the derived bounds.

A. Capacity of Gaussian MAC in the Finite Block-length

The works in [195]–[197] investigated simple bounds on the K-user MAC, where K is fixed and block length is finite. It was shown that, unlike the asymptotic case, OMA is strictly bounded below the capacity. It was further showed that the capacity can only be achieved with NOMA and joint decoding. In general, these bounds involve the evaluation of probabilities in $2K$-dimensional spaces, and thus can only be evaluated for small values of K. Similar results have been found in [198]–[200]. Other works aimed at investigating the penalty on the performance bounds introduced by the random activity of users and collisions [146]. This work showed that by considering collisions as interference, the throughput can be significantly increased. Moreover, it was also shown that there is a significant gap in performance between joint decoding and successive decoding.

B. The Gaussian Many Access Channel

The work in [201] introduced a new paradigm to this area of research, namely the many access paradigm. A case was considered where both the number of users and the block length go to infinity; however, the number of users can scale as fast as linearly with the block length. This assumption renders classical theory inapplicable, as the number of users can grow larger than the block length, but is typical for most, if not all, mMTC scenarios. When user activity is unknown at the receiver (random), the capacity is achievable by having transmitters concatenate a unique signature sequence to the payload and performing a two-stage decoding scheme at the receiver: (1) user detection, followed by (2) data decoding. The derived capacity for RA is the same as that with known access, minus a penalty factor characterized by the minimum length of the signature sequences needed to obtain an arbitrarily low and even vanishing error probability in active user detection. An error in active user detection can be either a missed detection or a false alarm. Similar to the results in [195], the authors in [201] concluded that successive decoding cannot achieve the sum capacity in this scenario unlike classical multiple access capacity regions. Overall, the derivations in [201] are heavily rooted in it being a CS problem. According to our knowledge, no attempts to date have shown to achieve or approach this capacity bound.

C. Random Coding Bound

Unlike the work in [201], the approach in [148] sought a capacity bound for the case where users utilize the same encoders/codebook (no unique signature sequences) which is more practical when the total number of users in the network is large. This is also known as symmetrical encoding. The sole task of recovering the unordered list of transmitted messages is assigned to the decoder, thus decoupling the role of user identification from data recovery, reasoning that the identification is part of the payload. Provided that this permutation invariance holds (i.e. the conditional channel output distributions are independent of the channel input permutations), [148] derives the random coding bound that dictates the limits on RA code lengths that can achieve an arbitrarily small error probability. The error probability is defined as the number of average fractions of correctly recovered messages. Thus, the error is defined on user level rather than a network level. For a K-user MAC, the RA codes should be such that the sum of any K or fewer unique codewords can be decoded reliably.

D. Achievability of the Random Coding Bound

Following on from the RA channel setup in [148], the work in [202] demonstrated the achievability using rateless codes. Unlike [201], where random user activity was shown to introduce a penalty on the capacity, [202] showed that the performance is the same in the first and second order, whether user activity is known or not. Moreover, for a symmetric multiple access channel, there is a single-threshold decoding rule rather than the more familiar $2K-1$ simultaneous threshold rules. The considered class of rateless codes differs slightly from that in the literature. Here, codes are not rateless in that codewords vary in length, but rather that decoding varies in time. Moreover, traditional rateless codes, such as Raptor codes, assume arbitrary decoding times and single-bit feedback to terminate transmission, the codes in [202] considered that a single-bit feedback is transmitted at every time step indicating whether to continue or terminate transmission. In [202], users listen at finite and predetermined set of times thus allowing the feedback rate to vanish as the block length grows. The decoding times are fixed to n_1, n_2, such that n_i denotes the time at which the decoder believes there are i active users. After every coding epoch n_i, the receiver sends a feedback bit that indicates whether it is ready or not to decode. Other works inspired by the random coding bound and the system model introduced by it can be found in [147], [149]. However, the work therein focuses on proposing practical and low complexity system architectures rather than achieving or approaching the capacity bound.

E. Future Directions in Information Theory for Grant-free NOMA

To this end, it is important to note that there is no satisfactory bound to date for grant-free NOMA in information theory. In fact, the work on RA coding bounds by information theorists is at risk of being disjoint from the practical designs and challenges being dealt with. Motivated by this, we end this section with a summarized road map representing how we envision the work in information theory on grant-free NOMA to evolve. We also shed light on how these bounds are driving design and feasibility. Firstly, the main future directions are summarized as follows.

- *Finite block-length approximations:* As aforementioned, to date, there is no satisfactory RA coding bound for the finite block-length regime [203], [204]. We expect the activities in this field to continue to refine their approximations and even aim for higher order approximations (with the capacity being the first order approximation, and the dispersion being the second order approximation) [205]. The reason behind this is that second order approximations cover block-lengths larger than 100 bits. However, we know from IoT applications, such as monitoring, that some messages can be as small as a few bits.
- *Channel models:* Most RA bounds at the moment consider the additive white Gaussian noise channel with few works on fading [206]. The same applies for multiple antenna transmissions [207].
- *User activity model:* We have witnessed the literature evolve in its assumption on user activity in grant-free access. Researchers have moved from the sparsity model often tied with CS techniques, to a K out of T model (K active users out of a total of T users). We suspect there is no right answer to this problem, especially with the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

heterogeneity of IoT applications. We also suspect there will be more to come as these applications become more and more of a reality.

- *Unsourced random coding:* The current trend with grant-free NOMA is to allow the users to reuse the same codebook. Moreover, the information theorists have deemed the identification of users at the MAC layer not required and consider the decoding task as the task of recovering an unordered set of transmitted messages. This is often referred to as ”unsourced random coding” in the literature. It will be interesting to see how this assumption plays out as the reuse of codebooks is notorious for being spectrally inefficient. On the other hand, even though the identification of users should be properly handled by the upper layers, there is usually metadata that is recovered at the MAC layer that is required for that, whose overhead could be currently underestimated.

Finally, we shed light on how the efforts in information theory are driving the feasibility aspect in this field, despite the gap between the existing RA methods and the derived bounds being still quite significant. We highlight three main metrics that are a culprit for this gap.

- *Low complexity:* The proposed T-fold ALOHA [148], [149], [205] to meet the random coding bound, still exhibits a significant gap. The gap was further reduced via sequential interference cancellation scheme in [208] through density evolution. Other efforts in this domain have experimented with CS as well to reduce the gap [209]. The urge for cheap sensors will continue to push this domain of work in the future.
- *Energy efficiency:* The transmission of multiple users over fading channels is known to significantly increase the minimum required energy per information bit. Recent works have demonstrated significant gains in energy efficiency using iterative receivers to mitigate this issue [206]. However, these gains along with the best achievable trade-offs have yet to be quantified analytically.
- *Asynchronous transmissions:* Most models so far in this realm assume frame and slot synchronized transmissions. This is based on the use of regularly spaced beacons. However, beacon-free transmissions to support frame-asynchronous scenarios will surely improve the energy efficiency of the system by reducing unnecessary overhead and processing power.

VIII. Practical Challenges and Solutions for Grant-free UL NOMA

NOMA has demonstrated significant potential to provide massive connectivity and large spectral gains. However, to support/optimize grant-free transmission, resource definition, allocation, and selection is important. Moreover, user synchronization, active user and data detection, potential collision management, HARQ, link adaptation, and power control procedures would need to be investigated. Some of these practical challenges are highlighted in Fig. 17, and are discussed as follows.

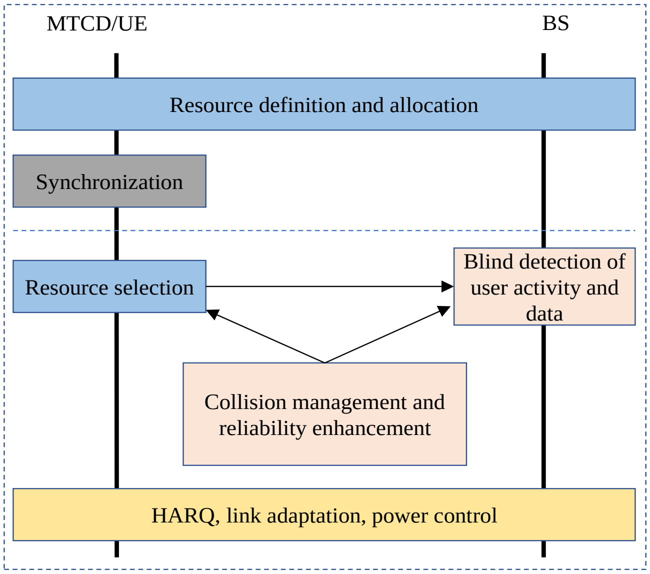

MTCD/UE
BS
Resource definition and allocation
Synchronization
Resource selection
Blind detection of user activity and data
Collision management and reliability enhancement
HARQ, link adaptation, power control

Fig. 17: Overview of technicalities in grant-free transmission.

A. Resource Definition, Allocation and Selection

In order to enable grant-free access, the radio resource for transmission should be defined before any grant-free transmission starts, and be known to both the UE and BS. The pre-defined resource for grant-free transmission can be similar to the CTU defined earlier in Fig. 10. Each CTU may include time-frequency resources, and may combine with a set of pilots for channel estimation and/or UE activity detection, and a set of MA signatures (e.g., codebooks/sequences/interleavers) for robust signal transmission and interference whitening, etc. Moreover, the size and location of time/frequency resources, as well as the pilot/signature patterns associated with it should be pre-defined, as shown earlier in Fig. 10. Once resources are defined, the resource allocation problem is to study how to allocate the CTUs to different UEs. It is possible for the BS to allocate the CTUs to users. However, to enable more autonomous transmission, the users can select a CTU i.e., some specific pilot and signature pattern. The selection of the specific pilot and signature can either be done randomly from the resource pool or according to some pre-defined rule.

B. Synchronization

Grant-free/contention-based UL transmission is expected to be supported by transmissions without close-loop time alignment signaling or a RA process (RACH-less grant-free). In such scenarios, if the timing offset between randomly transmitting MTCDs is larger than the cyclic prefix (CP), it is referred to as asynchronous transmission. This asynchronous transmission will cause a tremendous complexity increase for MTCD detection and decoding at the receiver side.

To aid in asynchronous transmissions, well-designed preambles can assist the active user identification, timing-offset/frequency-offset estimation, and channel estimation. Some preamble transmission procedures for the mMTC UL are provided in [210] and research is ongoing in this area.

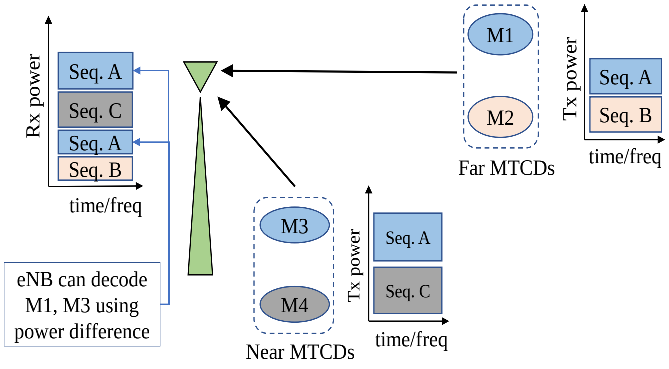


Fig. 18: Use of multiple MA signatures for collision handling.

Another possible solution is to attempt synchronization using information from the DL signal [211]. Given DL synchronization, a UE can adjust its UL transmit timing and, in many cases [212], achieve UL synchronization (time-offsets between UEs within CP length) without close-loop timing advance command. To avoid increasing the reception complexity and operation difficulty, time misalignment can be limited by using proper CP length and symbol duration.

C. Blind Activity Detection and Data Decoding

In grant-free transmission, since the BS has no prior information of when a MTCD may initiate transmission, it has to detect on each CTU which MTCDs have transmitted. Such MTCD activity detection is normally preferred to be done jointly with data decoding to reduce latency and overhead. This blind detection is crucial to the performance of grant-free UL communication. However, how to do the blind detection and based on what to detect is the major problem that needs to be investigated with this approach.

One option is to use pilots, as was done in grant-free UL SCMA [132]. In this case, pilots may serve the purpose of both the MTCD activity detection and channel estimation. In this context, efficient pilot designs for joint user activity detection and channel estimation should be studied.

Alternatively, some non-coherent detection techniques have recently been proposed in the literature for massive NOMA in grant-free access [213]–[215]. As well, the CS-MUD and ML techniques described earlier can also be used for efficient data recovery.

Finally, for many proposed MUD schemes, error propagation is also a major concern. For example, the SIC receiver may suffer from imperfections, and the SIC error is then propagated and effects other overlapped MTCDs detection. A comprehensive discussion and performance evaluation of different advanced MUD receivers for grant-free transmissions is provided in [216].

D. Collision Management and Reliability Enhancement

In grant-free transmissions, it is likely that more than one MTCD will select the same time-frequency resource. In the code/interleave domain NOMA, which includes spreading, interleaving or codebook-based signatures, colliding MTCDs are still separable through differences in their signatures or pattern vectors. However, if some MTCDs happen to use the same pattern vector, a hard collision is said to have taken place and these MTCDs will suffer from each other's signal interference. Hence, design of UL NOMA schemes should consider how to deal with collisions of NOMA signatures for grant-free/contention-based UL transmissions [217], and the impact of such collisions should be studied in each proposed scheme [218].

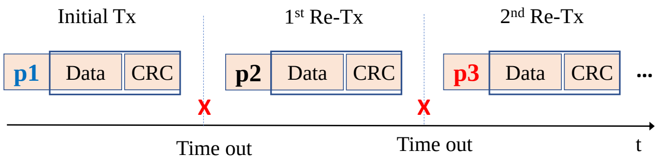


Fig. 19: Initial transmission and re-transmission of one user with mapped pilots.

Some possible solutions to reduce MA signature collisions include efficient MA signature design, larger resource pool, detection optimization, MA resource management, and mode switching mechanism (between grant-based and grant-free) [219]. In the context of resource pool expansion, Fig. 18 shows an example of integrating spreading based NOMA with PD-NOMA. Here 4 MTCDs transmit their data to the BS, where each MTCD randomly chooses a spreading sequence from a resource pool. However, due to random selection, MTCDs 1 and 3 choose the same signature sequence A. If the signals of these two MTCDs reach the BS with close received power, the BS will be unable to differentiate between them, and a collision is said to have taken place. As one of the countermeasures, PD-NOMA is introduced in the system, where users can choose between different available power levels, thereby adding another dimension. As a result, if the received signals of M1 and M3 at the BS have different received powers, the BS can easily decode their data despite of having same spreading sequence, thereby creating the possibility of spreading sequence reuse.

E. Link Adaptation and Power Control

Link adaptation represents the matching of the modulation, coding, and other signal transmission parameters to the conditions of the radio link. Efficient link adaptation results in better utilization of the network resources, detection error rate reduction, energy efficiency, reduced latency etc. Generally, link adaptation is based on the channel state information. However, in grant-free transmissions, MTCDs might not have the exact UL channel status.

One solution to link adaption is to consider the channel between MTCDs and BS to be reciprocal in each direction (a case in time division duplexing). Hence, the MTCDs can estimate their channel to the BS using the periodically received pilot/reference signals from BS, and correspondingly adjust their transmission parameters to facilitate data recovery at the BS. For instance, in the grant-free Random NOMA scheme explained earlier, the MTCDs after estimating their channel

TABLE V: Summary of practical challenges, their description, and possible solutions

Practical Challenge	Description	Possible Solutions
Resource definition, allocation, and selection	• How to define a grant-free resource? • How to allocate resources to users?	• CTU based resource; containing MA signature, pilots etc. • Predefined/pre-allocated CTU or randomly selected. • Random pilot and MA signature selection by MTCD.
UL synchronization	• RACH-less random transmission. • Large timing offsets cause synchronization issues. • Detection complexity/inefficiency of MUD.	• MTCD adjusts UL timing based on DL synchronization. • Well-designed preambles. • Proper CP length and/or symbol duration.
Blind detection	• BS has no prior information of active MTCDs. • Joint MTCD activity and data detection needed. • How and based on what?	• Use of pilot symbols for activity detection and channel estimation. • CTU design: may contain MCS and other MUD information. • CS and ML based activity/data detection.
Collision management	• How to minimize collisions? • How to manage once a collision happens?	• Efficient MA signature design and increasing resource pool. • Mixing multiple domains e.g., spreading and power. • Efficient ACK/NACK feedback and re-transmissions.
Link adaptation and power control	• How to choose MCS? • How to achieve power control?	• Different CTUs linked with different MCS and power values. • Use DL channel estimation for choosing UL MCS, power, etc.
HARQ	• How to know a failed transmission? • How to merge original and HARQ re-transmissions? • Distinguishing transmissions/re-transmissions at BS.	• Efficient design of ACK/NACK feedback to users. • Re-transmissions from a user with different pilot values.

can adjust their transmission power so that their received power at the BS is always the same fixed value, which helps the BS in load estimation over a sub-band [144].

Another solution is to divide the radio resources or CTUs into orthogonal MA blocks (MABs). Different MABs occupy different radio resources and may adopt different transmit parameter settings i.e., transmission block sizes, MCSs, transmit power, etc. Configurations of the MABs can be broadcasted by the BS. During the data transmission phase, any active MTCD first selects one MAB followed by MA signature or pattern vector. At the BS, MUD can be performed in parallel on each MAB. In addition, each MAB may be assigned a limited number of signatures, thereby reducing the computational complexity of blind detection at the receiver.

F. Hybrid Automatic Repeat Request (HARQ)

In UL grant-free transmissions, a user waits for a fixed time period to determine if its previous transmission is successful or if a re-transmission is needed. This is usually achieved through an ACK/NACK feedback from the BS. However, unlike grant-based transmission, the BS is not aware of which UEs are transmitting information in advance. Therefore, the BS needs to perform user activity and data detection, prior to the ACK/NACK feedback. In case of collisions, a NACK may never be sent since the attempted transmission is not detected and a time-out protocol at the MTCD is important. In either case, HARQ re-transmissions are of prime importance to guarantee the reliability of data. HARQ can efficiently merge new transmissions with the previous one. However, an issue is how to identify the first transmission and the re-transmissions for a HARQ process.

One potential solution to identify HARQ transmissions is that the BS can explicitly schedule re-transmissions via DL control signaling. Another efficient mechanism to address this problem is to use different pilots mapped to transmission and re-transmissions to identify the transmissions of a same packet by a particular user. An example of this is shown in Fig. 19, where a user does one transmission and two re-transmissions. Pilots p1, p2 and p3 are mapped on initial transmission, 1st re-transmission and 2nd re-transmission, respectively. If the BS successfully identified all pilots during the transmission, it can still potentially decode the signal by combing all the uncoded packets. Furthermore, due to the grant-free nature of the transmission, the HARQ procedure can be different from LTE scheduled HARQ [211]. A detailed insight into the HARQ process, and potential HARQ techniques for grant-free transmissions is provided in [220].

Based on the discussion in this section, the identified practical challenges, their description, and possible solutions are summarized in Table. V.

IX. Future Directions

For scheduling-based NOMA schemes, comprehensive analysis of various MA signature (spreading, interleaving, scrambling, multiple domains) based schemes exists in the literature. However, for grant-free/contention-based UL NOMA schemes, there is a need for further investigations, new designs, and integration of multiple technologies to deal with various challenges. In 3GPP RAN1 meetings, it was agreed that, parallel to new designs, the following should be continuously studied [221].

- Resource allocation/selection options; a) randomly by user, b) pre-configured by BS.
- UL synchronization (DL synchronization assumed) by considering two cases; timing offsets between users are within or greater than CP.

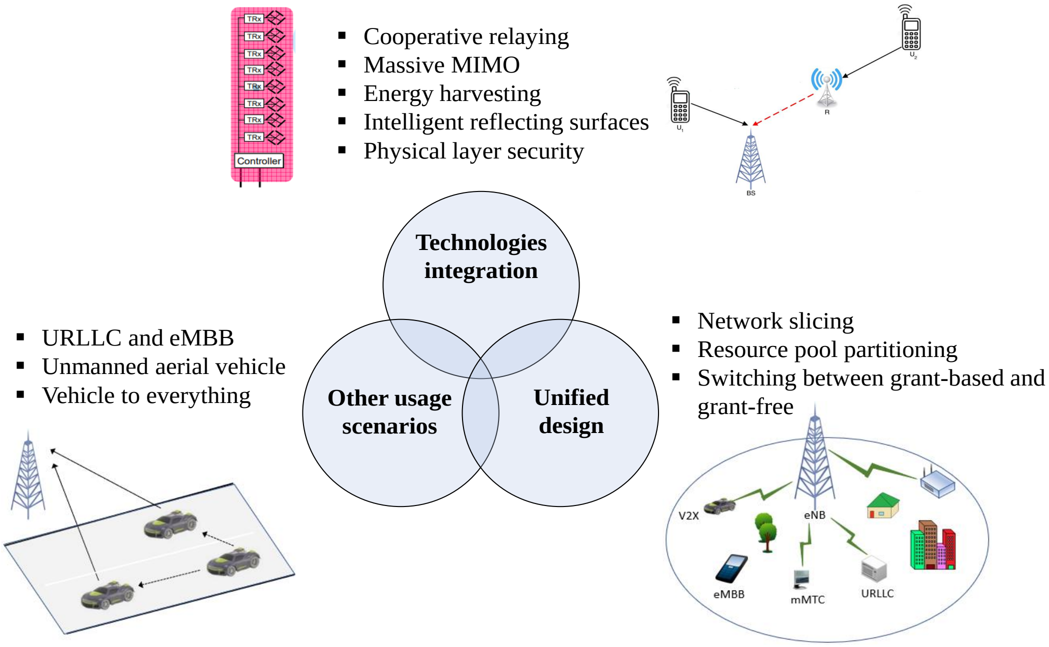


Fig. 20: Future research directions of grant-free NOMA.

- Collision handling of time/frequency resources and MA signatures (e.g., code, sequence, interleaver pattern).
- Re-transmissions/repetitions of failed transmissions and potential combining, e.g., HARQ.
- Link adaptation e.g., MCS/signature selection.
- Relationship between grant-free and grant-based transmissions and associated user behavior.
- Advanced receiver capabilities and complexity analysis.
- Requirement for power control.

These study items are a pursuit of solutions to the practical challenges explained earlier. In addition to these study items, other innovative options need to be explored for facilitating/improving grant-free communication in mMTC and other use cases. In this context, we discuss some other future directions to improve the performance of grant-free transmissions, as shown in Fig. 20.

A. Unified Framework for IoT Use Cases

As aforementioned, mMTC traffic is just one of the prominent IoT traffic categories besides eMBB (majorly HTC) and URLLC. Considering the coexistence of all these traffic types with extremely diverse QoS requirements, a unified framework is required, where each of these use cases can be supported using a single backbone system. Some potential directions to achieve these goals are highlighted in Fig. 20, and are explained as follows.

1) Network Slicing with Grant-free UL NOMA: Considering a variety of IoT use cases, and a diverse range of QoS requirements, using the concept of network slicing may provide additional flexibility of resource allocation to different use cases. The objective is to allow a physical mobile network operator to partition its network resources to allow for very different users, so-called tenants, to multiplex over a single physical infrastructure. The most commonly cited example in 5G discussions is sharing of a given physical network to simultaneously run mMTC, eMBB, and URLLC.

NOMA with network slicing has been under focus recently. In [222], vital challenges of resource management pertaining to network slicing using the NOMA-based scheme are highlighted. In this context, efficient solutions for resource management in network slicing for NOMA-based scheme are provided. Moreover, a slice-based virtual resource scheduling scheme with NOMA to enhance the QoS of the system is proposed in [223]. While network slicing with NOMA has been explored to some extent in the existing literature, these works only focus on grant-based access. Therefore, in order to provide a unified framework with grant-free/contention-based transmissions support, an unmet need is to devise grant-free UL NOMA based solutions using network slicing. An example scenario is further shown in Fig. 21.

2) Resource Pool Partitioning for Grant-free UL NOMA: Similar to network slicing, depending on applications/services, packet payload sizes from multiple users could be different. To allow efficient resource use for different payload sizes, and to reduce BS receiver complexity, multiple NOMA sub-regions can be defined within one NOMA resource pool, where each NOMA sub-region may be tailored for one particular MCS, transmission block size or coverage enhancement level if supported for mMTC [224]. Moreover, some OMA based regions can be defined/reserved for priority use cases.

3) Switching Between Grant-free and Grant-based: For a unified network-based operation, using dynamic grant to override grant-free transmissions can enable switching between grant-based and grant-free transmissions and provide flexibility to the scheduler for handling urgent events or reconfiguring resources. For example, users may be able to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2020.2996032, IEEE Communications Surveys & Tutorials

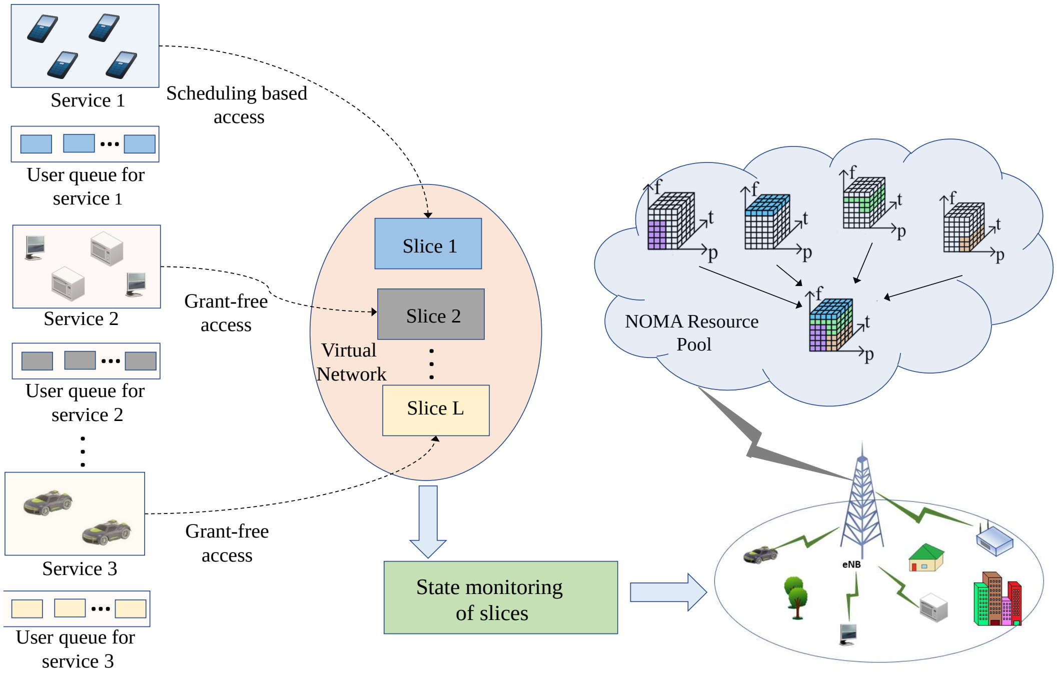


Fig. 21: Coexistence of grant-based and grant-free use cases through network slicing.

access different type of services, e.g., URLLC and eMBB, and switching between grant-based and grant-free may be useful in this context [211].

B. Integration of Grant-free NOMA with Other Technologies

The integration of scheduling/grant-based NOMA with other technologies is supported through various studies. However, when it comes to grant-free NOMA, there is less work found in the existing literature. Similar to grant-based NOMA, the performance of grant-free UL NOMA can be further improved by its integration with other cutting-edge technologies, as explained next.

1) Cooperative Communications: According to the European telecommunications standards institute, and developments by 3GPP on mMTC [6], [225], [226], three basic mMTC scenarios have been identified as follows:

- **Direct access:** A MTCD can access the BS without any intermediate device. This is also referred to as a direct 3GPP connection [226]. This is the simplest access method, and the one that has been considered in majority of the literature referred in this article. However, direct access may lead to traffic congestion and excessive signaling overhead when the number of MTCDs becomes very high.
- **Gateway access:** A MTCD can obtain cellular connectivity through a MTC gateway, which is a dedicated device for data relaying between the BS and a group of MTCDs, and does not generate its own traffic. This is also termed as an indirect 3GPP connection [226].
- **Coordinator access:** A group/cluster of MTCDs obtain cellular connectivity through a coordinator (temporary MTC gateway), which itself is also a MTCD with its own traffic.

Direct access is the point of focus in most of the existing literature as it is simple, but may lead to worse traffic congestion if the number of MTCDs is very high. Moreover, gateway and coordinator access are also critical, as group data transmission from a dedicated gateway or a temporary coordinator may reduce the overall power consumption of all the MTCDs and extend their service life, which is a key goal in mMTC/IoT. Some additional complex access scenarios might apply e.g., when a personal area network of MTCDs (a person wearing several smart wearables) attempts a gateway/coordinator access, or when the MTCDs and the relay/coordinator belong to different subscribers [226]. The network should also aim to provide service continuity for devices switching between various access types (e.g., gateway to direct access or vice versa), their continuous authorization, and flexibility of choosing a radio access technology (licensed or unlicensed) [226].

The poor channel quality of some distant-from-BS users is also a reason behind the importance of gateway and coordinator access. In this context, the relaying-based access can be further studied by considering half/full-duplex relaying mechanisms for grant-free UL NOMA transmissions. Furthermore, as the relays normally operate in either decode-and-forward or amplify-and-forward modes [227], their use in relaying-based grant-free UL NOMA can be further investigated.

2) Massive MIMO: Massive multiple-input multiple-output (MIMO) can significantly increase the spectral efficiency of wireless communication systems through aggressive spatial multiplexing [228]. It is known that massive MIMO-OMA systems, with prevalent linear processing at the BS, can achieve the best spectral efficiency in under-loaded conditions; i.e., the number of users many fewer than the number of antennas [229]. Therefore, such integration may not be able to support massive connectivity in overloaded systems, where the number of devices exceeds the number of antennas at the BS. To this end, the use of massive MIMO-NOMA has shown great potential to tackle the connectivity requirements of overloaded systems [230].

In massive MIMO-NOMA, the large number of antennas at the BS can be used to generate multiple beams for separating users in the space domain, creating the so-called spatial division multiple access (SDMA). Moreover, within each beam, contrary to MIMO-OMA, signals of multiple users are transmitted as a superimposed signal as in conventional NOMA, hence leading to the concept of intra-beam superposition transmission. At the receivers, a combination of interference cancellation approaches, e.g., spatial filtering and SIC, are applied to remove the inter-beam and intra-beam interference respectively. With proper design, the scheme can potentially capture the benefits of both massive MIMO and NOMA. In UL scenarios where large number of users transmit to a BS with many antennas, the BS can exploit the spatial diversity to perform MUD [231]–[233].

However, the design of massive MIMO-NOMA is by no means trivial. For instance, in conventional DL PD-NOMA, the BS allocates different power factors to the users by comparing their scalar channel gains. However, in massive MIMO-NOMA scenarios, the channel of each user is represented as a vector due to multiple transmit antennas at the BS. Hence, it becomes difficult to order the users in terms of their channel gains and decide their power allocations, especially when the number of users is large. This requires significant research in solving problems of user clustering, resource allocation, and receiver design for DL massive MIMO-NOMA systems.

Although there exists significant research work on DL massive MIMO-NOMA in literature, the UL scenario has got limited focus. Moreover, almost all of the existing work focuses on grant-based transmissions, where the user clusters and different NOMA related transmission parameters are predefined or pre-configured. Hence, the potential of massive MIMO-NOMA in supporting grant-free UL transmissions has not been exploited yet. Significant efforts in this context are needed to bridge the gap between grant-free transmissions and massive MIMO-NOMA for UL scenarios. Moreover, the existing massive MIMO-NOMA literature focuses on both, the sub-6 GHz and millimeter wave, frequency bands, which exhibit different channel characteristics. In this context, further research is required to design and analyze massive MIMO based grant-free NOMA schemes by considering the dynamics of these frequency bands. In addition to this, the existing massive MIMO-NOMA studies majorly rely on the spatial domain for pilot allocations, user clustering, etc. Considering the recent exploitation of angle domain with array signal processing techniques [234], angular models of the massive MIMO-NOMA channels can be leveraged for system design based on angle information.

3) Energy Harvesting: IoT devices are power-limited, and maintaining a long life of these devices is of prime importance. Energy harvesting offers the possibility of providing self-sufficient and self-sustaining means of communications; a major step towards realizing green communications. While energy is normally harvested from external natural resources, it could also be harvested from ambient electromagnetic signals. In this context, technologies such as simultaneous wireless information and power transfer (SWIPT) and wireless-powered communication networks (WPCN) have been thoroughly studied in existing literature. It is clear that such notions, when combined with NOMA, can realize massive connectivity while providing energy harvesting opportunities to the IoT devices for a long battery life [235]–[237].

While all these works are focused on grant-based NOMA schemes where power split ratios or time slots for information and energy transfer are predefined, the use of energy harvesting in grant-free UL NOMA poses significant research challenges. The sporadically transmitting mMTC devices need to have some kind of synchronization with the BS to know these power or time split parameters. Moreover, the existing literature on (grant-based) NOMA with energy harvesting considers centralized transmissions either from a BS to multiple NOMA users in DL or vice versa. However, as mMTC network can have multi-directional traffic, more general system models and energy harvesting mechanisms need to be analyzed. Moreover, the existing works focus on a repetitive phase wise procedure, where devices harvest energy and then transmit their data successfully in the information transfer phases. However, in grant-free information transmission, a device may experience collisions for many consecutive time slots, resulting in energy loss. In such scenarios, mechanisms to keep the devices efficiently energized are needed.

4) Reconfigurable Intelligent Surfaces: Reconfigurable intelligent surfaces (RIS, [238]), also known as large intelligent metasurfaces (LIM, [239]) and intelligent reflecting surfaces (IRS, [240]), have been considered as a potential technology for improving the energy and spectral efficiencies of wireless networks by altering the wireless environment. RISs are electromagnetic materials that reflect incident signals with a phase that is electronically controlled using integrated electronics. These surfaces enable the telecommunication operators to shape the electromagnetic response of the environmental objects that are distributed across the network. Hence, the propagation environment can be intentionally and deterministically controlled in order to improve the signal quality at the receiver.

Recently, integration of RISs with NOMA has gained research interest. The works in [241]–[243] focus on RIS aided PD-NOMA for DL transmissions; [241] considers a single-antenna BS to design the passive beamforming weights at the RISs for enhancing the spectral efficiency of users, [242] uses machine learning to design a novel framework for the deployment and passive beamforming of RISs, and [243] considers multi-antenna BS to jointly optimize the beamforming vectors at the BS and passive beamforming at the RIS for minimizing transmit power of the BS and enhancing spectral efficiency of users respectively.

While there exists some literature on the integration of RIS with NOMA, the work is still in its embryonic stage. These works mainly focus on DL scenarios and focus in particular on one NOMA scheme i.e., PD-NOMA. There is a clear reason for this as PD-NOMA specifically relies on the power difference of transmitted/received signals, and shaping/controlling the electromagnetic response of the environmental objects has a clear advantage for such a scheme. However, even for PD-NOMA, there is no research work to design the passive beamforming weights of RISs that can benefit the UL users' data transmission. In this context, a unified or bi-directional design to facilitate both DL and UL communication is needed. Moreover, considering the availability of a variety of NOMA schemes, the role of RISs in these schemes is still to be analyzed. Finally, all existing work, even on PD-NOMA, focuses on grant-based transmissions, where the beamforming weights of RISs are optimized by considering the location of BS and the pair/group of users. In this context, considering the sporadic nature of transmissions by devices in grant-free access, the integration with RIS is still an open research problem.

5) Physical Layer Security: Due to the broadcast nature of wireless transmissions, achieving secure communication in the presence of external/internal eavesdroppers is of prime importance. Physical security is a powerful tool for achieving provably unbreakable secure communications. These security schemes exploit various physical aspects of communication channels between the communicating entities to ensure secure data transmission. Most notable techniques in this context are artificial noise addition to cause confusion at the eavesdropper, beamforming approaches, transmit antenna selection, and relay-based physical security systems. These approaches, in general, are applicable to NOMA systems and have been investigated in the existing literature [244]–[246].

Secrecy of data is a critical performance bottleneck in NOMA schemes where a number of users are multiplexed over the same radio resources. Most of the physical layer security solutions for NOMA in literature are limited to two-user DL PD-NOMA scenarios. A generalization of these solutions to large-scale networks with multiple users will be of great importance. Moreover, all these works focus on grant-based schemes, which may not be directly applicable in grant-free transmissions where there is minimal control signaling between the transmitting and receiving nodes. This situation is further exacerbated by the fact that mMTC or general IoT traffic is distributed and multi-directional, which shrinks the window for secrecy control through mechanisms popular in conventional cellular communication. Hence, there is a significant gap in research in this area to apply physical layer security solutions to grant-free access.

C. Application of Grant-free NOMA to Other IoT Service Types and Use Cases

The existing grant-free NOMA schemes majorly focus on mMTC scenarios. However, as aforementioned, there also exist some other prominent service types and use cases of IoT, where there may be potential benefits of grant-free NOMA. For instance, while most of the grant-free NOMA work focuses on mMTC traffic, the URLLC services can also be a target application area. It is understandable that grant-free NOMA can reduce the access delay, which is crucial to the low-latency requirements of URLLC. However, there exists a couple of issues especially related to the data reliability aspect. Firstly, mutual interference due to the non-orthogonal transmissions and the corresponding error rate issues need to be properly evaluated to suit the high-reliability constraints of URLLC [138]. Moreover, the grant-free NOMA agreed for mMTC is still contention-based, which means that the probability of collisions with other traffic is still there. In this context, preserving UL resources for URLLC traffic followed by grant-free transmission (contention-free access) is one solution, but may not be efficient since traffic occasions for URLLC cannot be anticipated [128], [211].

Considering these issues, the potential of grant-free NOMA to be applied for these different service types is still an open research area. Moreover, as 3GPP agreed that NOMA should also be considered for diversified usage scenarios besides mMTC, there exists some initial work on the use of grant-free NOMA for URLLC. For instance, [247] analyzes the error rate performance of grant-free NOMA for URLLC by using SCMA scheme. Similarly, some discussions on the use of grant-free access for URLLC are provided in [248], [249]. The basic proposal is to use OMA-based grant-free transmissions for URLLC as baseline, which can later be extended to the NOMA implementation. In this context, latency and error rate should be properly analyzed to see the potential usage scenarios of grant-free NOMA for URLLC traffic. Moreover, considering the variety of grant-free NOMA schemes available in literature, the suitable schemes especially in terms of low error rate need to be considered for URLLC use cases. Hence, the research on grant-free NOMA for URLLC still contains a great potential.

In the eMBB service type, the traffic is dominantly DL, and the main target goals are higher system capacity and achievable data rates. To achieve these goals, tight control and heavy signaling are needed. Hence, the use of grant-based NOMA is practical, and has therefore gained significant research interest as evident through majority of the existing literature and the surveys listed in Tab. II. Even for the UL of eMBB, the overhead of control signaling is much less than the transmission data size, and is not a major issue. Hence, the use of grant-free NOMA schemes, particularly suitable for sporadic short packet UL traffic, are less likely to be necessary for eMBB services [138].

While most of the research on grant-free NOMA considers conventional network scenarios, there are some other use cases of IoT with different system/channel conditions. The potential of grant-free NOMA to be employed in such use cases is discussed next.

1) Unmanned Aerial Vehicle: Unmanned aerial vehicles (UAVs) have in recent years shown significant potential to fill a variety of roles in numerous applications. They can be deployed as flying BSs to leverage the strength of line-of-sight connections and effectively enhance the coverage and connectivity of wireless communication systems in exceptional situations e.g., sports events, concerts, disasters, military situations, etc [250]. In the same scenarios, UAVs can also be used as ordinary users connected to the ground BSs for live coverage and monitoring purposes. To this end, the majority of the existing literature focuses on the former scenario of UAV-BS-enabled communication systems [251].

The work in [252] aims at rate maximization for a DL NOMA system where a single UAV-BS serves multiple ground users, [253] presents UAV-centric and user-centric communication strategies for DL scenarios where multiple UAVs serve ground users through PD-NOMA, and [254], [255] discuss the use of UAVs as relays in DL/UL cooperative PD-NOMA to improve the communication between BS and far users.

The majority of the UAV-NOMA work focuses on scheduling-based NOMA where the goal is to optimize various parameters such as UAV placement, trajectory design, beam-width control, power allocation etc. However, there still remains a lack of literature on how the devices can access the channel resources using grant-based or grant-free procedures for UL transmissions; an initial work on NOMA-based RA for UL of UAV-enabled communications is presented in [256]. This is critical due to a number of reasons. Firstly, UAV-BSs may need to simultaneously serve a variety of users with different service requirements e.g., in disaster or military situations, where some tasks are delay intolerant and cannot rely on grant-based access. For instance, the work in [257] discusses beam-width design for DL of UAV-enabled URLLC; but not UL communication. Secondly, even in some of these situations, the UAV may be serving just as a relay, which may pose further delays if grant acquisition is required. Moreover, as multiple UAVs can also be used to cooperatively serve a large group of users at a particular location, the serving UAV selection by a user and channel access may also incur delays. Significant research in this area is therefore needed to analyze various channel access mechanisms in UAV-enabled communications. Moreover, generalized application scenarios such as UAV-to-everything (U2X) or air-to-everything (A2X) need to be studied for a more unified UAV-enabled communication framework.

2) Vehicle-to-everything: Recent years have seen a boom in the possible applications of vehicular networks for intelligent transportation such as driver assistance, active safety, and traffic efficiency. In this context, vehicle-to-everything (V2X) provides an integrated system or framework of vehicular networking to enable vehicle communication with each other and beyond [258]. It encompasses three types of communications i.e., vehicle-to-vehicle (V2V), vehicle-to-pedestrian (V2P), and vehicle-to-infrastructure/network (V2I/N), referring to the communication between a vehicle and a roadside unit/network. With a huge number of vehicles in the system transmitting/receiving a variety of short packet data (which may or may not have tolerance for delay and reliability), V2X exhibits properties of both mMTC and URLLC use cases of IoT. Although IEEE 802.11p provides security and upper layer specifications for V2X services, its unpredictable latency and limited transport capacity limit its ability to achieve the expected performance. In this context, LTE is considered as a promising solution for supporting V2X services as it can achieve large cell coverage, controllable latency, and high data rates even in high mobility scenarios [259]. However, due to the use of OMA in LTE, and a large number of vehicles in a dense network, congestion and access delays are inevitable.

To this end, the use of NOMA for V2X services is frequently discussed in literature [260]. Capable of achieving highly overloaded transmissions over limited resources, NOMA provides a new dimension for V2X services to alleviate the resource collision and access delay. The application and performance analysis of NOMA in various V2X scenarios has been considered in [261]–[263]. Though the work is still in its early stages, given the diverse types of V2X communications, major challenges in the line of work and future research directions are summarized below.

In the OMA-based V2X systems, semi-persistent scheduling is applied in which the resources are reserved or booked by the vehicles every few transmission periods. While this ensures the URLLC aspect of V2X for some vehicles, providing massive connectivity to all vehicles due to limited orthogonal resources may cause them huge access delays. The use of NOMA with this semi-persistent scheduling is one possible way to satisfy the massive connectivity requirements of V2X communications. However, considering multiple vehicles loaded over the same radio resources, their high mobility, and the fact that NOMA receivers need prior knowledge about the real time channel state information of users to perform efficient MUD, a new scheduling scheme combining dynamic power control with semi-persistent scheduling needs to be considered. To further reduce the access delay, grant-free contention-based NOMA transmissions for V2X communication is another open issue. While it can significantly reduce latency, the data reliability becomes a major point of concern as the efficiency of MUD in NOMA depends a lot on prior knowledge. In the viewpoint of all this, a possible future direction of research is to carefully categorize the various types of V2X signaling based on latency and reliability requirements. Correspondingly, grant-based and grant-free resource pool partitions can be made for vehicles to choose from and transmit their data according to the type of signal to be transmitted. Moreover, some other issues, such as NOMA signature selection and contention-based backoff mechanisms in case of collisions, are also possible areas of future research.

X. Conclusion

It is agreed that 5G should support grant-free/contention-based UL transmissions for mMTC. To enable such transmissions, the use of NOMA has gained significant research interest from academia and industry. In this context, unlike the existing works, this article provides a comprehensive survey of NOMA from a grant-free connectivity perspective. The article starts by establishing a case for unconventional solutions to tackle the huge and sporadic mMTC traffic. The concept of NOMA is then introduced, where illustration of a general transmitter and different possibilities of NOMA signature design are discussed. From this point onward, the major focus of the survey is on the grant-free access and the role of NOMA to enable this. Different grant-free NOMA schemes proposed by academia and industry are categorized into four different types, and are thoroughly discussed. The article later provides an information theoretic perspective of grant-free NOMA schemes, and highlights the gaps between theory and practice. In the end, practical challenges and possible future directions are discussed.

The survey explains that the grant-free versions of various NOMA schemes differ in the fact that the data transmission by a user is sporadic in an arrive and go manner, resource selection can be randomly done by the users, and an important consideration is the need for efficient receiver design to identify the active devices and decode their transmitted data. While there exists a variety of works in this context, the survey explains that the work is still somewhat unstructured. There seems to be some ambiguity about which services and use cases of the IoT family can benefit from grant-free NOMA schemes. While most of the grant-free NOMA works focus on the mMTC use case, exploring its potential benefits in other IoT use cases is also crucial. It is expected that the work in this survey will provide comprehensive information regarding grant-free NOMA, its potential to be employed in different use cases of IoT especially mMTC, the major practical challenges, and possible future directions.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart. 2015.

[2] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[3] L. D. Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Trans. Ind. Inform.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[4] S. Lien, K. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[5] ITU-R, "IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond," Int. Telecommun. Union, Geneva, Switzerland, ITU Recommendation M.2083-0, Sep. 2015.

[6] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 12–18, Jun. 2014.

[7] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 184–192, Jul. 2012.

[8] Ericsson, "Cellular networks for massive IoT-enabling low power wide area applications," Ericsson, Stockholm, Sweden, Technical Report Uen 284 23-3278, Jan. 2016.

[9] 3GPP, "Study on communication services for critical medical applications (release 17)," 3GPP, Valbonne, France, Technical Report 22.826 V1.0.0, May 2019.

[10] ——, "Study on enhancement of 3GPP support for 5G V2X services (release 16)," 3GPP, Valbonne, France, Technical Report 22.886 V16.2.0, Dec. 2018.

[11] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: From mobile to embedded internet," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 36–43, Apr. 2011.

[12] P. Jain, P. Hedman, and H. Zisimopoulos, "Machine type communications in 3GPP systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 28–35, Nov. 2012.

[13] ETSI, "Machine-to-machine communications (M2M); functional architecture," ETSI, Sophia Antipolis, France, Technical Specification TS 102 690 V2.1.1, Oct. 2013.

[14] Cisco, "Cisco annual internet report (2018–2023)," White Paper, Mar. 2020.

[15] Ericsson, "Ericsson mobility report," Ericsson, Stockholm, Sweden, Technical Report EAB-18:004510Uen, Revision A, Jun. 2018.

[16] 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 Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

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

[18] Y. Wang, B. Ren, S. Sun, S. Kang, and X. Yue, "Analysis of non-orthogonal multiple access for 5G," *China Commun.*, vol. 13, no. 2, pp. 52–66, Feb. 2016.

[19] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart. 2017.

[20] S. Yang, P. Chen, L. Liang, J. Zhu, and X. She, "Uplink multiple access schemes for 5G: A survey," *ZTE Commun.*, vol. 15, no. S1, pp. 31–40, Jun. 2017.

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

[22] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[23] M. Basharat, W. Ejaz, M. Naeem, A. M. Khattak, and A. Anpalagan, "A survey and taxonomy on nonorthogonal multiple-access schemes for 5G networks," *Trans. Emerging Telecommun. Technol.*, vol. 29, no. 1, p. e3202, Jan. 2018.

[24] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 629–646, 1st Quart. 2018.

[25] Z. Ding, M. Xu, Y. Chen, M. Peng, and H. V. Poor, "Embracing non-orthogonal multiple access in future wireless networks," *Frontiers Inf. Technol. Electronic Engineering*, vol. 19, no. 3, pp. 322–339, May 2018.

[26] N. Ye, H. Han, L. Zhao, and A.-h. Wang, "Uplink nonorthogonal multiple access technologies toward 5G: a survey," *Wireless Commun. Mobile Comput.*, vol. 2018, Jun. 2018.

[27] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart. 2018.

[28] M. Mohammadkarimi, M. A. Raza, and O. A. Dobre, "Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 40–50, Dec. 2018.

[29] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA schemes for enabling technologies of future wireless networks," *arXiv preprint arXiv:1909.08011*, Sep. 2019.

[30] M. Vaezi, G. A. Aruma Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cognitive Commun. Netw.*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

[31] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, "Probabilistic rateless multiple access for machine-to-machine communication," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6815–6826, Dec. 2015.

[32] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, Dec. 2018.

[33] S. Sesia, I. Toufik, and M. Baker, *LTE-the UMTS long term evolution: From theory to practice*. Hoboken, NJ, USA: John Wiley & Sons, 2009.

[34] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G evolution: HSPA and LTE for mobile broadband*. New York, NY, USA: Academic press, 2008.

[35] N. Khan, J. Mišić, and V. Mišić, "Priority-based machine-to-machine overlay network over LTE for a smart city," *J. Sensor Actuator Netw.*, vol. 7, no. 3, p. 27, Jul. 2018.

[36] D. T. Wiriaatmadja and K. W. Choi, "Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 33–46, Jan. 2015.

[37] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.

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

[39] S. Lien, T. Liau, C. Kao, and K. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 27–32, Jan. 2012.

[40] J. Cheng, C. Lee, and T. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Houston, TX, USA, Dec. 2011, pp. 368–372.

[41] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[42] 3GPP, "Study on RAN improvements for machine type communications," 3GPP, Valbonne, France, Technical Report 37.868 V11.0, Sep. 2011.

[43] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-advanced random-access mechanism for massive machine-to-machine (M2M) communications," in *Proc. 27th World Wireless Res. Forum (WWRF) Meeting*, Oct. 2011, pp. 1–5.

[44] 3GPP, "Evolved universal terrestrial radio access (E-UTRA): Physical channels and modulation," 3GPP, Valbonne, France, Technical Report 36.211 V14.11.0, Jun. 2019.

[45] Ericsson, "NR PRACH preamble design," 3GPP TSG-RAN WG1 Meeting #87, Reno, NV, USA, document R1-1611904, Nov. 2016.

[46] ZTE, ZTE Microelectronics, "Considerations on the preamble design for grant-free non-orthogonal MA," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608955, Oct. 2016.

[47] Huawei, HiSilicon, "Evaluation on CP types for UL transmissions," 3GPP TSG-RAN WG1 Meeting #87, Reno, NV, USA, document R1-1611198, Nov. 2016.

[48] Ericsson, "NR random-access response design," 3GPP TSG-RAN WG1 Meeting #87, Reno, NV, USA, document R1-1611911, Nov. 2016.

[49] Qualcomm Incorporated, "RACH timeline considerations," 3GPP TSG-RAN WG1 Meeting #87, Reno, NV, USA, document R1-1612035, Nov. 2016.

[50] ZTE Corporation, ZTE Microelectronics, "On 2-step random access procedure," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608969, Oct. 2016.

[51] Nokia, Alcatel-Lucent Shanghai Bell, "Random access principles for new radio," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609737, Oct. 2016.

[52] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[53] P. Wang, J. Xiao, and L. P, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, Sep. 2006.

[54] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[55] NTT DOCOMO, "5G radio access: Requirements, concept and technologies," White Paper, Jul. 2014.

[56] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Cham, Switzerland: Springer, 2019.

[57] MCC Support, "Final report of 3GPP TSG RAN WG1 #84bis v1.0.0," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-165448, May 2016.

[58] ——, "Final report of 3GPP TSG RAN WG1 #86 v1.0.0," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608562, Oct. 2016.

[59] 3GPP, "WF on common features and general framework of MA schemes," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1610956, Oct. 2016.

[60] China Telecom, "Classification of candidate UL non-orthogonal MA schemes," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-167445, Aug. 2016.

[61] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Naha, Japan, Nov. 2013, pp. 770–774.

[62] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, Mar. 2015.

[63] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials," in *Proc. Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, Marrakech, Morocco, Oct. 2015, pp. 1–6.

[64] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Tokyo, Japan, Oct. 2015, pp. 18–23.

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

[66] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[67] M. B. Shahab, M. Irfan, M. F. Kader, and S. Young Shin, "User pairing schemes for capacity maximization in non-orthogonal multiple access systems," *Wireless Commun. Mobile Comput.*, vol. 16, no. 17, pp. 2884–2894, Dec. 2016.

[68] M. B. Shahab, M. F. Kader, and S. Y. Shin, "On the power allocation of non-orthogonal multiple access for 5G wireless networks," in *Proc. Int. Conf. Open Source Syst. Technol. (ICOSST)*, Lahore, Pakistan, Dec. 2016, pp. 89–94.

[69] M. B. Shahab and S. Y. Shin, "User pairing and power allocation for non-orthogonal multiple access: Capacity maximization under data reliability constraints," *Physical Communication*, vol. 30, pp. 132–144, Oct. 2018.

[70] M. B. Shahab, M. F. Kader, and S. Y. Shin, "A virtual user pairing scheme to optimally utilize the spectrum of unpaired users in non-orthogonal multiple access," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1766–1770, Dec. 2016.

[71] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 458–461, Mar. 2016.

[72] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 781–785.

[73] Z. Sheng, X. Su, and X. Zhang, "A novel power allocation method for non-orthogonal multiple access in cellular uplink network," in *Proc. Int. Conf. Intell. Environments (IE)*, Seoul, South Korea, Aug. 2017, pp. 157–159.

[74] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.

[75] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[76] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[77] M. B. Shahab and S. Y. Shin, "Time shared half/full-duplex cooperative NOMA with clustered cell edge users," *IEEE Commun. Lett.*, vol. 22, no. 9, pp. 1794–1797, Sep. 2018.

[78] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient joint user-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5119–5131, Jun. 2019.

[79] F. Fang, Z. Ding, W. Liang, and H. Zhang, "Optimal energy efficient power allocation with user fairness for uplink MC-NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1133–1136, Aug. 2019.

[80] M. Zeng, N. Nguyen, O. A. Dobre, Z. Ding, and H. V. Poor, "Spectral- and energy-efficient resource allocation for multi-carrier uplink NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9293–9296, Sep. 2019.

[81] M. W. Baidas, M. S. Bahbahani, E. Alsusa, K. A. Hamdi, and Z. Ding, "Joint D2D group association and channel assignment in uplink multi-cell NOMA networks: A matching-theoretic approach," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8771–8785, Dec. 2019.

[82] Y. Sun, Z. Ding, X. Dai, and O. A. Dobre, "On the performance of network NOMA in uplink CoMP systems: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5084–5098, Jul. 2019.

[83] L. Zhang, W. Li, Y. Wu, X. Wang, S. Park, H. M. Kim, J. Lee, P. Angueira, and J. Montalban, "Layered-division-multiplexing: Theory and practice," *IEEE Trans. Broadcasting*, vol. 62, no. 1, pp. 216–232, Mar. 2016.

[84] 3GPP, "Study on downlink multiuser superposition transmission (MUST) for LTE (release 13)," 3GPP, Valbonne, France, Technical Report 36.859 V13.0.0, Dec. 2015.

[85] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[86] D. Guo and C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, Apr. 2008.

[87] J. van de Beek and B. M. Popovic, "Multiple access with low-density signatures," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Nov. 2009, pp. 1–6.

[88] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *Proc. IEEE 71st Veh. Technol. Conf. (VTC)*, Taipei, Taiwan, May 2010, pp. 1–5.

[89] M. Al-Imari, M. A. Imran, R. Tafazolli, and D. Chen, "Subcarrier and power allocation for LDS-OFDM system," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC)*, Yokohama, Japan, May 2011, pp. 1–5.

[90] L. Wen, "Non-orthogonal multiple access schemes for future cellular systems." Ph.D. dissertation, University of Surrey, Guildford, U.K, 2016.

[91] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Annu. Int. Symp. Personal, Indoor, Mobile Radio Commun. (PIMRC)*, London, UK, Sep. 2013, pp. 332–336.

[92] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, "Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 4782–4787.

[93] Huawei, HiSilicon, "Sparse code multiple access (SCMA) for 5G radio transmission," 3GPP TSG-RAN WG1 Meeting #84b, Busan, Korea, document R1-162155, Apr. 2016.

[94] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC)*, Vancouver, BC, Canada, Sep. 2014, pp. 1–5.

[95] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, UK, Jun. 2015, pp. 2918–2923.

[96] Y. Du, B. Dong, Z. Chen, J. Fang, and X. Wang, "A fast convergence multiuser detection scheme for uplink SCMA systems," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 388–391, Aug. 2016.

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

[98] A. Bayesteh, H. Nikopour, M. Taherzadeh, H. Baligh, and J. Ma, "Low complexity techniques for SCMA detection," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[99] X. Dai, S. Chen, S. Sun, S. Kang, Y. Wang, Z. Shen, and J. Xu, "Successive interference cancelation amenable multiple access (SAMA) for future wireless communications," in *Proc. IEEE Int. Conf. Commun. Syst.*, Macau, China, Nov. 2014, pp. 222–226.

[100] CATT, "Candidate solution for new multiple access," 3GPP TSG-RAN WG1 Meeting #84b, Busan, Korea, document R1-163383, Apr. 2016.

[101] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (PDMA) for cellular future radio access," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nanjing, China, Oct. 2015, pp. 1–5.

[102] 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 Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, Apr. 2017.

[103] X. Dai, Z. Zhang, B. Bai, S. Chen, and S. Sun, "Pattern division multiple access: A new multiple access technology for 5G," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 54–60, Apr. 2018.

[104] B. Ren, X. Yue, W. Tang, Y. Wang, S. Kang, X. Dai, and S. Sun, "Advanced IDD receiver for PDMA uplink system," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–6.

[105] Fujitsu, "Initial LLS results for UL non-orthogonal multiple access," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164329, May 2016.

[106] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for internet of things," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC)*, Nanjing, China, May 2016, pp. 1–5.

[107] LG Electronics, "Considerations on DL/UL multiple access for NR," 3GPP TSG-RAN WG1 Meeting #84b, Busan, Korea, document R1-162517, Apr. 2016.

[108] H. Hu and J. Wu, "New constructions of codebooks nearly meeting the Welch bound with equality," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1348–1355, Feb. 2014.

[109] Nokia, Alcatel-Lucent Shanghai Bell, "Non-orthogonal multiple access for new radio," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-165019, May 2016.

[110] Intel Corporation, "Multiple access schemes for new radio interface," 3GPP TSG-RAN WG1 Meeting #84b, Busan, Korea, document R1-162385, Apr. 2016.

[111] MediaTek Inc, "New uplink non-orthogonal multiple access schemes for NR," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-167535, Aug. 2016.

[112] Qualcomm Incorporated, "RSMA," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164688, May 2016.

[113] ——, "Candidate NR multiple access schemes," 3GPP TSG-RAN WG1 Meeting #84b, Busan, Korea, document R1-163510, Apr. 2016.

[114] ETRI, "Low code rate and signature based multiple access scheme for new radio," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164869, May 2016.

[115] Li Ping, Lihai Liu, Keying Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, Apr. 2006.

[116] Nokia, Alcatel-Lucent Shanghai Bell, "Performance of interleave division multiple access (IDMA) in combination with OFDM family waveforms," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-165021, May 2016.

[117] Samsung, "Non-orthogonal multiple access candidate for NR," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-163992, May 2016.

[118] S. V. Bana and P. Varaiya, "Space division multiple access (SDMA) for robust ad hoc vehicle communication networks," in *Proc. IEEE Intell. Transportation Syst. Conf. (ITSC)*, Oakland, CA, USA, Aug. 2001, pp. 962–967.

[119] D. Fang, Y. Huang, Z. Ding, G. Geraci, S. Shieh, and H. Claussen, "Lattice partition multiple access: A new method of downlink non-orthogonal multiuser transmissions," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[120] M. A. Naim, J. P. Fonseka, and E. M. Dowling, "A building block approach for designing multilevel coding schemes," *IEEE Commun. Lett.*, vol. 19, no. 1, pp. 2–5, Jan. 2015.

[121] NTT DOCOMO, "Uplink multiple access schemes for NR," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-165174, May 2016.

[122] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive multiple access based on superposition raptor codes for cellular M2M communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 307–319, Jan. 2017.

[123] Intel Corporation, "Grant-less and non-orthogonal UL transmissions in NR," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-167698, Aug. 2016.

[124] Huawei, HiSilicon, "Discussion on grant-free transmission," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-166095, Aug. 2016.

[125] ZTE, "Grant-based and grant-free multiple access for mMTC," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164268, May 2016.

[126] Lenovo, "Uplink grant-free access for 5G mMTC," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609398, Oct. 2016.

[127] CATT, "Consideration on grant-free transmission," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608757, Oct. 2016.

[128] DOCOMO, "Discussion on multiple access for UL mMTC," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-167392, Aug. 2016.

[129] E. Balevi, F. T. A. Rabee, and R. D. Gitlin, "ALOHA-NOMA for massive machine-to-machine IoT communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–5.

[130] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.

[131] M. Elkourdi, A. Mazin, E. Balevi, and R. D. Gitlin, "Enabling slotted Aloha-NOMA for massive M2M communication in IoT networks," in *Proc. IEEE 19th Wireless Microwave Technol. Conf. (WAMICON)*, Sand Key, FL, USA, Apr. 2018, pp. 1–4.

[132] K. Au, L. Zhang, H. Nikopour, E. Yi, A. Bayesteh, U. Vilaipornsawai, J. Ma, and P. Zhu, "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Austin, TX, USA, Dec. 2014, pp. 900–905.

[133] ZTE, "Contention-based non-orthogonal multiple access for UL mMTC," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164269, May 2016.

[134] Nokia, Alcatel Lucent Shanghai Bell, "Basic principles of contention-based access," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-167252, Aug. 2016.

[135] Qualcomm, "RSMA and SCMA comparison," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-166358, Aug. 2016.

[136] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 853–857.

[137] Huawei, HiSilicon, "LLS results for uplink multiple access," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164037, May 2016.

[138] ZTE, "Discussion on multiple access for new radio interface," 3GPP TSG-RAN WG1 Meeting #84b, Busan, Korea, document R1-162226, Apr. 2016.

[139] ZTE, ZTE Microelectronics, "Discussion on grant-free concept for UL mMTC," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-166405, Aug. 2016.

[140] ZTE, "System level performance evaluation for MUSA," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608952, Oct. 2016.

[141] ——, "Link-level performance evaluation for MUSA," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608953, Oct. 2016.

[142] ——, "Receiver implementation for MUSA," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-164270, May 2016.

[143] ——, "Receiver details and link performance for MUSA," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-166404, May 2016.

[144] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sep. 2017.

[145] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[146] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[147] ——, "A multi-layer grant-free NOMA scheme for short packet transmissions," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[148] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2523–2527.

[149] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access gaussian channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2528–2532.

[150] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[151] I. Bar-David, E. Plotnik, and R. Rom, "Forward collision resolution-a technique for random multiple-access to the adder channel," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1671–1675, Sep. 1993.

[152] T. Yang, L. Yang, Y. J. Guo, and J. Yuan, "A non-orthogonal multiple-access scheme using reliable physical-layer network coding and cascade-computation decoding," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1633–1645, Mar. 2017.

[153] L. Yang, T. Yang, Y. Xie, J. Yuan, and J. An, "Multiuser decoding scheme for K-user fading multiple-access channel based on physical-layer network coding," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1046–1049, May 2016.

[154] J. Goseling, M. Gastpar, and J. H. Weber, "Random access with physical-layer network coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3670–3681, Jul. 2015.

[155] J. Hong, W. Choi, and B. D. Rao, "Sparsity controlled random multiple access with compressed sensing," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 998–1010, Feb. 2015.

[156] F. Fazel, M. Fazel, and M. Stojanovic, "Random access compressed sensing over fading and noisy communication channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2114–2125, May 2013.

[157] F. Monsees, C. Bockelmann, and A. Dekorsy, "Reliable activity detection for massive machine to machine communication via multiple measurement vector compressed sensing," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Austin, TX, USA, Dec. 2014, pp. 1057–1062.

[158] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communication," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC)*, Glasgow, UK, May 2015, pp. 1–5.

[159] A. T. Abebe and C. G. Kang, "Compressive sensing-based random access with multiple-sequence spreading for MTC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[160] M. Alam and Q. Zhang, "A survey: Non-orthogonal multiple access with compressed sensing multiuser detection for mMTC," *arXiv preprint arXiv:1810.05422*, Oct. 2018.

[161] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC)*, Boston, MA, USA, Sep. 2015, pp. 1–5.

[162] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal*, vol. 26, no. 3, pp. 301–321, May 2009.

[163] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, Jul. 2016.

[164] J. Tan, W. Ding, F. Yang, C. Pan, and J. Song, "Compressive sensing based time-frequency joint non-orthogonal multiple access," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcasting (BMSB)*, Nara, Japan, Jun. 2016, pp. 1–4.

[165] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 972–974, Jul. 2012.

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

[167] N. Vaswani and J. Zhan, "Recursive recovery of sparse signal sequences from compressive measurements: A review," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3523–3549, Jul. 2016.

[168] 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. 19 983–19 993, Sep. 2017.

[169] J. Liu, G. Wu, S. Li, and O. Tirkkonen, "Blind detection of uplink grant-free SCMA with unknown user sparsity," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[170] A. T. Abebe and C. G. Kang, "Comprehensive grant-free random access for massive low latency communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[171] 3GPP, "Study on non-orthogonal multiple access (NOMA) for NR (release 16)," 3GPP, Valbonne, France, Technical Report 38.812 V0.2.0, Nov. 2018.

[172] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions," *IEEE Access*, vol. 7, pp. 137 184–137 206, Sep. 2019.

[173] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proc. Int. Conf. Eng. Appl. Neural Netw.* Springer, Sep. 2016, pp. 213–226.

[174] T. J. O'Shea, K. Karra, and T. C. Clancy, "Learning to communicate: Channel auto-encoders, domain specific regularizers, and attention," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Limassol, Cyprus, Dec. 2016, pp. 223–228.

[175] M. Borgerding and P. Schniter, "Onsager-corrected deep learning for sparse linear inverse problems," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Washington, DC, USA, Dec. 2016, pp. 227–231.

[176] N. Samuel, T. Diskin, and A. Wiesel, "Deep MIMO detection," in *Proc. IEEE 18th Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[177] Y. Jeon, S. Hong, and N. Lee, "Blind detection for MIMO systems with low-resolution ADCs using supervised learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[178] S. Dörner, S. Cammerer, J. Hoydis, and S. t. Brink, "Deep learning based communication over the air," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132–143, Feb. 2018.

[179] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, Feb. 2019.

[180] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[181] M. Kim, N. Kim, W. Lee, and D. Cho, "Deep learning-aided SCMA," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 720–723, Apr. 2018.

[182] D. A. Awan, R. L. G. Cavalcante, M. Yukawa, and S. Stanczak, "Detection for 5G-NOMA: An online adaptive machine learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[183] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA-based heterogeneous IoT with imperfect SIC," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2885–2894, Apr. 2019.

[184] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020.

[185] J. Luo, J. Tang, D. K. C. So, G. Chen, K. Cumanan, and J. A. Chambers, "A deep learning-based approach to power minimization in multi-carrier NOMA with SWIPT," *IEEE Access*, vol. 7, pp. 17 450–17 460, Jan. 2019.

[186] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, Nov. 2018.

[187] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, Dec. 2018.

[188] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile internet of things," *IEEE Trans. Ind. Inform.*, vol. 15, no. 5, pp. 2995–3005, May 2019.

[189] N. Ye, X. Li, H. Yu, L. Zhao, W. Liu, and X. Hou, "DeepNOMA: A unified framework for NOMA using deep multi-task learning," *IEEE Trans. Wireless Commun.*, pp. 1–1, Jan. 2020.

[190] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569–3582, Jul. 2019.

[191] W. Kim, Y. Ahn, and B. Shim, "Deep neural network based active user detection for grant-free NOMA systems," *IEEE Trans. Commun.*, pp. 1–1, Jan. 2020.

[192] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631–9640, Oct. 2018.

[193] R. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 2, pp. 124–142, Mar. 1985.

[194] T. M. Cover and J. A. Thomas, *Elements of information theory*. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[195] E. MolavianJazi and J. N. Laneman, "A second-order achievable rate region for gaussian multi-access channels via a central limit theorem for functions," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6719–6733, Dec. 2015.

[196] ——, "A random coding approach to gaussian multiple access channels with finite blocklength," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2012, pp. 286–293.

[197] E. M. Jazi and J. N. Laneman, "Simpler achievable rate regions for multiaccess with finite blocklength," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 36–40.

[198] L. V. Truong and V. Y. F. Tan, "On the gaussian MAC with stop-feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2303–2307.

[199] Y. Huang and P. Moulin, "Finite blocklength coding for multiple access channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 831–835.

[200] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Second-order rate region of constant-composition codes for the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 157–172, Jan. 2015.

[201] X. Chen, T. Chen, and D. Guo, "Capacity of gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516–3539, Jun. 2017.

[202] M. Effros, V. Kostina, and R. C. Yavas, "Random access channel coding in the finite blocklength regime," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 1261–1265.

[203] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *arXiv preprint arXiv:1910.12678*, Oct. 2019.

[204] R. C. Yavas, V. Kostina, and M. Effros, "Gaussian multiple and random access in the finite blocklength regime," *arXiv preprint arXiv:2001.03867*, Jan. 2020.

[205] I. Zadik, Y. Polyanskiy, and C. Thrampoulidis, "Improved bounds on gaussian MAC and sparse regression via gaussian inequalities," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 430–434.

[206] S. S. Kowshik, K. Andreev, A. Frolov, and Y. Polyanskiy, "Energy efficient coded random access for the wireless uplink," *arXiv preprint arXiv:1907.09448*, Jul. 2019.

[207] W. Cao, A. Dytso, Y. Shkel, G. Feng, and H. V. Poor, "Sum-capacity of the MIMO many-access gaussian noise channel," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5419–5433, Aug. 2019.

[208] A. Glebov, N. Matveev, K. Andreev, A. Frolov, and A. Turlikov, "Achievability bounds for t-fold irregular repetition slotted ALOHA scheme in the gaussian MAC," in *Proc. IEEE Wireless Commun. Netw, Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[209] A. Pradhan, V. Amalladinne, A. Vem, K. R. Narayanan, and J.-F. Chamberland, "A joint graph based coding scheme for the unsourced random access gaussian channel," *arXiv preprint arXiv:1906.05410*, Jul. 2019.

[210] Nokia, Alcatel-Lucent Shanghai Bell, "Preamble transmission procedures for the mMTC uplink," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609653, Oct. 2016.

[211] Samsung, "Discussion on grant-free/contention-based non-orthogonal multiple access," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-166752, Aug. 2016.

[212] ——, "Non-orthogonal multiple access candidate for NR," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-163992, May 2016.

[213] H. Chen, Z. Dong, and B. Vucetic, "Noncoherent and non-orthogonal massive SIMO for critical industrial IoT communications," in *Proc. IEEE Int. Conf. Ind. Cyber Physical Syst. (ICPS)*, Taipei, Taiwan, May 2019, pp. 436–441.

[214] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[215] K. Senel and E. G. Larsson, "Joint user activity and non-coherent data detection in mMTC-enabled massive MIMO using machine learning algorithms," in *Proc. 22nd Int. ITG Workshop Smart Antennas (WSA)*, Bochum, Germany, Mar. 2018, pp. 1–6.

[216] Huawei, HiSilicon, "Advanced multi-user detectors for grant-free transmissions," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608855, Oct. 2016.

[217] Sony, "Non-orthogonal multiple access for uplink," 3GPP TSG-RAN WG1 Meeting #86, Gothenburg, Sweden, document R1-166651, Aug. 2016.

[218] Nokia, Alcatel-Lucent Shanghai Bell, "Collision handling for grant-free," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609648, Oct. 2016.

[219] Huawei, HiSilicon, "Solutions for collisions of MA signatures," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608860, Oct. 2016.

[220] ——, "The retransmission and HARQ schemes for grant-free," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608859, Oct. 2016.

[221] ZTE, "Discussion on grant-free transmission based on sensing," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609801, Oct. 2016.

[222] S. A. Kazmi, L. U. Khan, N. H. Tran, and C. S. Hong, "Network slicing: Radio resource allocation using non-orthogonal multiple access," in *Network Slicing for 5G and Beyond Networks*. Cham, Switzerland: Springer, May 2019, pp. 69–89.

[223] L. Tang, Q. Tan, Y. Shi, C. Wang, and Q. Chen, "Adaptive virtual resource allocation in 5G network slicing using constrained markov decision process," *IEEE Access*, vol. 6, pp. 61 184–61 195, Oct. 2018.

[224] Intel, "Grant-free UL transmissions in NR," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609499, Oct. 2016.

[225] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 184–192, Jul. 2012.

[226] 3GPP, "Feasibility study on new services and markets technology enablers for massive internet of things (release 14)," 3GPP, Valbonne, France, Technical Report 22.861 V14.1.0, Sep. 2016.

[227] Y. Zhang, Z. Yang, Y. Feng, and S. Yan, "Performance analysis of a novel uplink cooperative NOMA system with full-duplex relaying," *IET Commun.*, vol. 12, no. 19, pp. 2408–2417, Nov. 2018.

[228] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[229] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[230] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, Jun. 2019.

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

[232] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3000–3004, Mar. 2019.

[233] W. Hao, M. Zeng, G. Sun, O. Muta, O. A. Dobre, S. Yang, and H. Gacanin, "Codebook-based max–min energy-efficient resource allocation for uplink mmWave MIMO-NOMA systems," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8303–8314, Dec. 2019.

[234] J. Zhao, F. Gao, W. Jia, S. Zhang, S. Jin, and H. Lin, "Angle domain hybrid precoding and channel tracking for millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6868–6880, Oct 2017.

[235] Z. Ni, Z. Chen, Q. Zhang, and C. Zhou, "Analysis of RF energy harvesting in uplink-NOMA IoT-based network," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC)*, Honolulu, HI, USA, USA, Sep. 2019, pp. 1–5.

[236] Y. Ye, Y. Li, D. Wang, and G. Lu, "Power splitting protocol design for the cooperative NOMA with SWIPT," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–5.

[237] T. A. Zewde and M. C. Gursoy, "NOMA-based energy-efficient wireless powered communications," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 679–692, Sep. 2018.

[238] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[239] Z. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[240] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[241] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, L. Hanzo *et al.*, "Reconfigurable intelligent surface aided NOMA networks," *arXiv preprint arXiv:1912.10044*, Dec. 2019.

[242] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *arXiv preprint arXiv:2001.10363*, Jan. 2020.

[243] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in multi-antenna aided NOMA systems," *arXiv preprint arXiv:1910.13636*, Oct. 2019.

[244] M. Zeng, N. Nguyen, O. A. Dobre, and H. V. Poor, "Securing downlink massive MIMO-NOMA networks with artificial noise," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 685–699, Jun. 2019.

[245] H. Lei, J. Zhang, K. Park, P. Xu, I. S. Ansari, G. Pan, B. Alomair, and M. Alouini, "On secure NOMA systems with transmit antenna selection schemes," *IEEE Access*, vol. 5, pp. 17 450–17 464, Aug. 2017.

[246] A. Arafa, W. Shin, M. Vaezi, and H. V. Poor, "Securing downlink non-orthogonal multiple access systems by trusted relays," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[247] Huawei, HiSilicon, "Grant-free non-orthogonal MA for uplink URLLC," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1608869, Oct. 2016.

[248] CATT, "UL grant-free transmission for URLLC," 3GPP TSG-RAN WG1 Meeting #87, Reno, USA, document R1-1611398, Nov. 2016.

[249] CATR, "Discussion on non-orthogonal multiple access for URLLC usage scenario," 3GPP TSG-RAN WG1 Meeting #86b, Lisbon, Portugal, document R1-1609580, Oct. 2016.

[250] I. Bor-Yaliniz and H. Yanikomeroglu, "The new frontier in RAN heterogeneity: Multi-tier drone-cells," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 48–55, Nov. 2016.

[251] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, Feb. 2019.

[252] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "UAV-enabled communication using NOMA," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5126–5138, Jul. 2019.

[253] T. Hou, Y. Liu, Z. Song, X. Sun, and Y. Chen, "Exploiting NOMA for UAV communications in large-scale cellular networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6897–6911, Oct. 2019.

[254] X. Jiang, Z. Wu, Z. Yin, Z. Yang, and N. Zhao, "Power consumption minimization of UAV relay in NOMA networks," *IEEE Wireless Commun. Lett.*, pp. 1–1, Jan. 2020.

[255] W. Mei and R. Zhang, "Uplink cooperative NOMA for cellular-connected UAV," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 644–656, Jun. 2019.

[256] J. Seo, S. Pack, and H. Jin, "Uplink NOMA random access for UAV-assisted communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8289–8293, Aug. 2019.

[257] A. Han, T. Lv, and X. Zhang, "UAV beamwidth design for ultra-reliable and low-latency communications with NOMA," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2019, pp. 1–6.

[258] S. Sun, J. Hu, Y. Peng, X. Pan, L. Zhao, and J. Fang, "Support for vehicle-to-everything services based on LTE," *IEEE Wireless Commun.*, vol. 23, no. 3, pp. 4–8, Jun. 2016.

[259] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: a survey," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, May 2013.

[260] B. Di, L. Song, Y. Li, and Z. Han, "V2X meets NOMA: Non-orthogonal multiple access for 5G-enabled vehicular networks," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 14–21, Dec. 2017.

[261] B. Di, L. Song, Y. Li, and G. Y. Li, "Non-orthogonal multiple access for high-reliable and low-latency V2X communications in 5G systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2383–2397, Oct. 2017.

[262] L. P. Qian, Y. Wu, H. Zhou, and X. Shen, "Dynamic cell association for non-orthogonal multiple-access V2S networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2342–2356, Oct. 2017.

[263] D. Zhang, Y. Liu, L. Dai, A. K. Bashir, A. Nallanathan, and B. Shim, "Performance analysis of FD-NOMA-based decentralized V2X systems," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5024–5036, Jul. 2019.

Muhammad Basit Shahab received his BS in Electrical Engineering from University of Engineering and Technology (UET) Lahore, Pakistan, in 2009, followed by MS in Electrical Engineering from University of Management and Technology (UMT) Lahore, Pakistan, in 2011. He received his PhD from the Department of IT Convergence Engineering, Kumoh National Institute of Technology (KIT), South Korea, in February 2019. Currently, he is working as a postdoctoral research fellow at the School of Electrical Engineering and Computing, The University of Newcastle (UoN), Australia. His main research areas are non-orthogonal multiple access (NOMA), grant-free communications, internet of things (IoT), and cooperative communication. He received the best researcher of the year awards in 2017 and 2019 for Brain Korea 21 (BK21) Plus project, and the best thesis award, in his PhD.

Rana Abbas received the M.E. in 2013 and the Ph.D. degree in 2018, both in electrical engineering, from The University of Sydney, Australia. She is currently a researcher at the Centre of IoT and Telecommunications at The University of Sydney. Her research interests include channel coding, random access, machine type communications and IoT. She is the recipient of the Australian Postgraduate Awards Scholarship and the Norman I. Price scholarship from the Centre of Excellence in Telecommunications, School of Electrical and Information Engineering, The University of Sydney. She is also the winner of the Best Paper Award at IEEE PIMRC, 2018.

Mahyar Shirvanimoghaddam received his BSc and MSc degrees, both in Electrical Engineering, with 1st Class Honor in 2008 and 2010, respectively from Sharif University of Technology and University of Tehran, Iran. He received his PhD in Electrical Engineering (Telecommunications) from The University of Sydney, Australia, in 2015. He is currently a Lecturer at Centre for IoT and Telecommunications, The University of Sydney. His general research interests include Coding and Information Theory and Internet of Things technologies. Mahyar is an IEEE Senior Member and a Fellow of the Higher Education Academy (FHEA). He is the exemplary reviewer of IEEE Transactions on Communications (2016 and 2019) and IEEE Communication Letters (2016). He was selected as one of the Top 50 Young Scientists in the World by the World Economic Forum in 2018 for his contributions to the development on IoT technologies.

Sarah J. Johnson received the B.E. degree in electrical engineering and the Ph.D. degree from The University of Newcastle, Australia, in 2000 and 2004, respectively. She held a postdoctoral fellowship with NICTA, Australia's Information and Communications Technology Research Center of Excellence, an Australian Research Council Postdoctoral Research Fellowship and an Australian Research Council Future Fellowship. Sarah is currently a Professor of Electrical Engineering and Deputy Head of Faculty at The University of Newcastle, Australia. Her research interests include channel coding, machine type communications and IoT.