Towards the Standardization of Grant-free Operation and the Associated NOMA Strategies in the 3GPP

Ali Cagatay Cirik*, *Member, IEEE*, Naveen Mysore Balasubramanya*, *Member, IEEE*, Lutz Lampe, *Senior Member, IEEE*, Gustav Vos, *Member, IEEE* and Steve Bennett

***Abstract*—The dramatic increase in Internet traffic to and from wireless devices poses significant challenges for network operators. While the current growth of traffic is mostly due to consumers communicating more frequently and larger amounts of data over the wireless infrastructure, much of the future growth is predicted to originate from non-human operated devices or the so-called Internet of Things (IoT) communication. The third generation partnership project (3GPP) standardization activities towards the fifth-generation (5G) cellular systems envisages two IoT-centric scenarios - massive machine type communications (mMTC) and ultra reliable low latency communications (URLLC). In this article, we quantify the advantages of grant-free operation for 5G mMTC using latency, signaling overhead and power consumption aspects. We propose the high-level design of a new grant-free state of operation and explain its interaction with the legacy long term evolution (LTE) operating states. We also describe promising grant-free NOMA solutions with respect to synchronization and hybrid automatic repeat request (HARQ), and resource collision handling procedures. Our discussion and proposal relate closely to the current 5G standardization activities in the 3GPP and highlight solutions that can be easily integrated to the current LTE framework.**

***Keywords*—Grant-free, long term evolution (LTE), 5G New Radio (NR), non-orthogonal multiple access (NOMA).**

I. Introduction

Recent advances in wireless communications have provided a great stimulus and an essential foundation for efficiently supporting the Internet of Things (IoT). A primary advancement in this regard is the machine-to-machine (M2M) or machine type communications (MTC) technology, which drives a large variety of application domains. The high growth potential of MTC is a strong incentive for cellular wireless technology providers to participate in this market. However, supporting the IoT over cellular networks presents a new set of challenges, such as, handling massive growth in traffic, supporting diverse quality-of-service (QoS) requirements and reducing the energy consumption.

The current cellular technologies like the long term evolution (LTE) governed by the third generation partnership project (3GPP), have mainly focused on increasing the spectral efficiency of human operated devices. However, 3GPP quickly recognized the need for enhancing its standards to more efficiently support IoT applications and started adding M2M-related amendments from Release 10. Further, 3GPP is supporting the fifth-generation (5G) evolution through the 5G New Radio (NR) standard and 5G core network architecture design, which not only define the protocols and procedures in different layers, but also explore novel access mechanisms for IoT, such as, non-orthogonal multiple access (NOMA) and grant-free access.

A grant-free access mechanism enables the user equipment (UE) to transmit data in an arrive-and-go manner in the next available slot. Unlike the current grant-based access mechanism in LTE uplink (illustrated in Fig. 1), the UE using grant-free transmission need not wait for a specific uplink grant from the eNB. Such a scheme is more desirable for the two broad IoT use cases in 5G, namely massive machine type communications (mMTC) and ultra-reliable and low latency communications (URLLC), as it has the following advantages: 1) reduced transmission latency, 2) smaller signaling overhead due to the simplification of the scheduling procedure, and 3) improved energy efficiency (battery life) of the UEs with the reduction in signaling and the ON time of the UE. Whether grant-free mechanisms can satisfy the stringent requirements of URLLC ($<$ 0.5 ms latency with 99.9999% reliability) is still open for study in 5G. But the potential benefits of grant-free NOMA for uplink mMTC, which are more delay tolerant and have lower reliability requirements, have been demonstrated in [1]. Hence we focus on standardization activities grant-free mechanisms for 5G mMTC.

Generally, in grant-free access, due to the lack of UE scheduling on orthogonal time-frequency resources, there is a high probability that different UEs randomly choose the same resource blocks for the uplink transmission, resulting in the superposition of data of multiple UEs (collision). Therefore, an access mechanism like NOMA, which is both spectrally efficient and robust towards user superposition, is desirable. NOMA tackles the problem of multiuser superposition using either power-domain or code-domain multiplexing [2]. The synergy of NOMA and grant-free transmission leads to the following advantages for mMTC scenarios:

- *Reduced energy consumption:* Since grant-free NOMA

The first two authors have contributed equally to this work.

Ali Cagatay Cirik is with Ofinno Technologies, Reston, VA, 20190, USA (email: acirik@ofinno.com).

Naveen Mysore Balasubramanya is with the Department of Electrical Engineering, Indian Institute of Technology Dharwad, Karnataka, India (email: naveenmb@iitdh.ac.in).

Lutz Lampe is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada (email: lampe@ece.ubc.ca).

Gus Vos and Steve Bennett are with Sierra Wireless Inc., Richmond, BC, V6V 3A4, Canada (email: {gvos, sbennett}@sierrawireless.com).

This work was supported by Sierra Wireless Inc. and the Natural Sciences and Engineering Research Council of Canada (NSERC).

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

can support more users than grant-free OMA [2], the users can quickly obtain network access, thereby reducing the latency. With the improved robustness towards UE superposition as the number of UEs increases, NOMA also reduces the probability of retransmission. These improvements reduce the ON time of the devices and hence reduce their energy consumption[1], which is beneficial for battery constrained devices in 5G mMTC.

- *Enabling flexible service multiplexing:* Although OMA dynamically schedules heterogeneous users/services, it splits the available bandwidth for different services by allocating orthogonal time-frequency resources, which may lead to poor spectral efficiency. However, NOMA may improve the performance of heterogeneous services by allowing sharing of time-frequency resources. For instance, enhanced mobile broadband (eMBB) services can be multiplexed with delay tolerant, low data rate mMTC services using NOMA, such that the QoS requirements of both the services are satisfied.

NOMA mechanisms involving power allocation, code design, fairness analysis, user pairing, etc. have been investigated in e.g. [2]–[4]. However, the standardization activities towards a grant-free NOMA in 5G are still in a nascent stage. In this article, we provide insights into the 3GPP standardization activities for grant-free operation and the NOMA strategies corresponding to such an operation, with respect to the 5G NR standards.

- We focus on the key performance indicators of the mMTC scenario associated with the IoT - latency, signaling overhead and energy consumption.
- We discuss promising grant-free solutions for uplink synchronization, present our proposal for a new operating state for grant-free transmission and elaborate on its interactions with the legacy operating states in LTE.
- We describe the challenges and prospective solutions for enhanced hybrid automatic repeat request (HARQ) and resource collision handling in grant-free NOMA, which are being considered by the 3GPP for standardization.
- We also discuss some recent results for grant-free NOMA in terms of collision handling and multiuser detection (MUD).

The proximity of our work to the recent standardization activities provides a different perspective from those covered in previous literature [2]–[4]. Also, the featured solutions require minimal changes to the current standards, ensuring smooth integration to the current LTE framework.

II. Quantifying the Key Performance Indicators Driving Grant-Free Operation in the 3GPP

The current LTE system uses two categories of scheduling in order to transmit data packets: i) semi-persistent scheduling (SPS) and ii) dynamic scheduling. Although SPS reduces the signaling overhead and is efficient in periodic traffic such as voice over Internet protocol (VoIP), it is not suitable for scheduling infrequent small packets due to the bursty nature of traffic. On the other hand, dynamic scheduling is widely used for a variety of applications in eMBB. However, it requires the UE to obtain downlink synchronization, decoding of the system frame number and the system information blocks, followed by a random access channel (RACH) procedure to initiate a data transfer, as shown in Fig. 1. The RACH procedure uses four messages (Msg-1 to Msg-4) for preamble transmission and radio resource control (RRC) connection set-up with contention resolution. The earliest phase for uplink data transmission is at Msg-5, after the reception of the uplink grant from the base-station (eNB) (See Fig. 1).

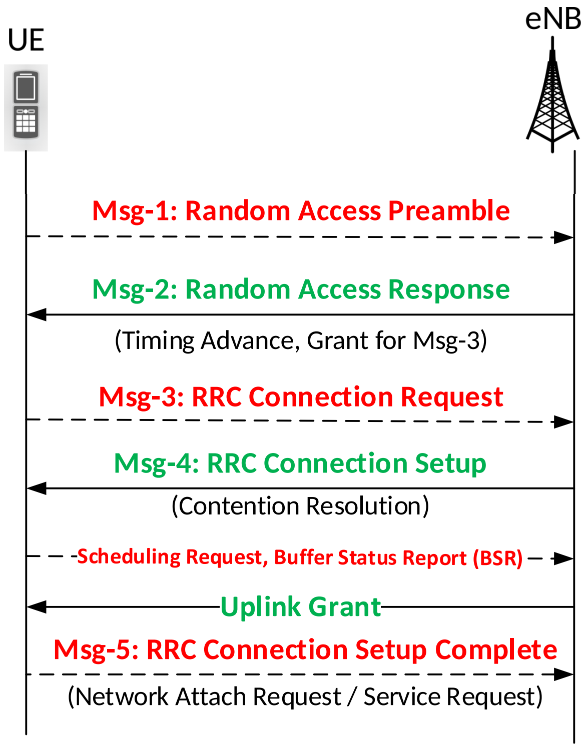

UE
eNB
Msg-1: Random Access Preamble
Msg-2: Random Access Response
(Timing Advance, Grant for Msg-3)
Msg-3: RRC Connection Request
Msg-4: RRC Connection Setup
(Contention Resolution)
Scheduling Request, Buffer Status Report (BSR)
Uplink Grant
Msg-5: RRC Connection Setup Complete
(Network Attach Request / Service Request)

Fig. 1. Random access and RRC procedures in LTE.

Accounting for the RACH procedure and subsequent message exchanges, we quantify the latency, signaling overhead and power consumption, which are the key performance indicators driving grant-free operation.

A. Latency

In the current LTE systems, UEs transmit a scheduling request (SR) and buffer status report (BSR) to ask eNB for an uplink grant (see in Fig. 1)[2]. In order to send an SR, a UE must wait for an SR opportunity on the physical uplink control channel (PUCCH) resource and upon receiving the SR, the eNB sends an uplink grant to the corresponding UE on the physical downlink control channel (PDCCH). The uplink

[1] The amount of reduced energy consumption will be quantified in Section II-C.

[2] LTE also allows a mechanism where the SR and the BSR can be transmitted as a part of Msg-3 during RRC set-up for the first transmission and then sent separately for subsequent uplink transmissions.

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

TABLE I. UPLINK LATENCY

Steps	Grant-based (TTI)	Grant-free (TTI)
Minimum waiting time for SR (in grant-based) or transmission opportunity (in grant-free)	1	0
UE sends SR on PUCCH	1	NA
eNB decodes SR and generates the scheduling grant	3	NA
Transmission of scheduling grant	1	NA
UE processing delay	3	NA
Transmission of uplink data and BSR	1	1
Data decoding in eNB and ACK/NACK generation	3	3
Total	13	4

1 TTI = 1 ms in LTE or 1 OFDMA symbol in 5G NR.
1 OFDMA symbol = 71.354 μs with 15 kHz subcarrier spacing and normal CP.
NA: Not applicable

TABLE II. SIGNALING OVERHEAD AND ENERGY CONSUMPTION

(a) Signaling overhead

Uplink Data size and CP size	40 bytes, Extended CP
Number of uplink data subcarriers	2 PRB pairs = 2 × 120 subcarriers = 240
Time taken for uplink data transmission	Case 1: 1 ms (2 PRB pairs in one subframe) Case 2: 2 ms (1 PRB pair in one subframe)
PRACH preamble duration	800 μs
Preamble transmission overhead (based on time taken)	Case 1: $\frac{800}{1000}$ = 80% Case 2: $\frac{800}{2000}$ = 40%
RAR and RRC connection set-up overhead (based on data size)	Message size = 51 bytes Overhead = $\frac{51}{40}$ = 127.5%
PDCCH for uplink grant overhead (based on subcarriers)	Grant size using DCI Format 0 = 36 CCEs = 36 subcarriers Overhead = $\frac{36}{240}$ = 15%

(b) Energy consumption share

Settings	Data transmission	Four-step RACH	Others
PL = 144 dB, RRC = 300 ms	75%	19%	6%
PL = 144 dB, RRC = 100 ms	82%	12%	6%
PL = 135 dB, RRC = 300 ms	36%	50%	14%
PL = 135 dB, RRC = 100 ms	48%	33%	19%

grant allows the UE to transmit its data on scheduled resources over physical uplink shared channel (PUSCH) [5]. In Table I, we provide details of the latency comparison between grant-free and grant-based scheduling in terms of the transmit time interval (TTI). For 1 TTI = 1 OFDMA symbol (as in 5G NR), the latency for the grant-based and grant-free transmissions can be calculated as 13×71.354 μs $= 0.927$ ms and 4×71.354 μs = 0.285 ms, respectively.

In [6], *fast uplink* transmission, where a UE is configured with a periodic uplink grant at every TTI, which eliminates SR messages and reservations of resources, is discussed as a way to reduce the latency of the first uplink transmission. However, periodic uplink grant requires reserved uplink resources on each TTI, which becomes difficult to get as the TTI size gets smaller. It is shown by simulations in [6] that the gains from using *fast uplink* access solutions are not significant with shorter TTI lengths, e.g. 1-symbol TTI used in 5G NR. Therefore, a grant-free transmission offering low latency is desirable, since it also reduces ON time and hence the energy consumption.

B. Signaling Overhead

In addition to the access delay, signaling overhead, which is associated with the massive number of UEs and infrequent small packets transmission, also hinders the usage of grant-based transmission. For example, mMTC applications, which are predicted to support massive connection densities (e.g., up to 10^6 per km^2), increase the total number of SRs sent from UEs and grants sent from the eNB. Moreover, mMTC is typically characterized with infrequent small packets. Therefore, the ratio of the number of bits required for signaling to that for data packets increases as the packet size decreases, resulting in a per-packet signaling overhead.

In Table II(a), we provide a quantitative analysis of the extra signaling overhead incurred from the grant-based transmission for mMTC applications with infrequent traffic and small packets. The extra signaling overhead for grant-based over grant-free scheme includes: preamble transmission, random access response (RAR), RRC connection set-up, UE SR/BSR in uplink and the control information of the uplink grant transmitted by the eNB using the PDCCH [1].

From Table II(a), it can be seen that grant-based access will incur a signaling overhead of 40% (or 80%), 127.5% and 15% for preamble transmission, RAR/RRC connection set-up and PDCCH for uplink grant, respectively, when compared to the "data-only" grant-free transmission. Hence, reduced signaling is also a major impetus for grant-free operation.

C. Energy Consumption

The long scheduling process in LTE involving message transmissions/exchanges in the uplink and downlink channels results in high energy consumption for the UEs with infrequent small packets. In addition, the UE chooses an RACH preamble in Msg-1 from a limited set of available uplink preamble patterns in LTE. Therefore, as the number of UEs increase, the probability of UEs choosing the same preamble increases, leading to failed transmissions. This initiates retransmissions, increasing the ON time of the UE and thus increasing the power consumption. As an example, we evaluate the energy consumption of mMTC devices during RACH procedure under different path-loss (PL) scenarios. Here, a discharge-unconstrained battery is modeled as having an energy capacity of 5 Wh at 3.6 V. The power consumption model is adopted from [7], where the "Transmit", "Receive", "Sleep", and "Standby (Hibernate)" operating modes have power consumption of 60 mW, 60 mW, 3 mW, and 0.015 mW, respectively. The scenarios where UEs transmit 200 bytes every 2 hours [7] under different PL and average RRC connection times are evaluated, and the energy consumption share for different processes are provided in Table II(b).

The column "Others" in Table II(b) refers to the sum energy consumption of the downlink synchronization, system information acquisition, sleep mode and transitions between hibernate and wake-up. As seen in Table II(b), a significant portion of the energy is consumed during the four-step RACH procedure. Thus eliminating the RACH procedure through

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

grant-free transmissions is beneficial to reduce the UE power consumption for mMTC.

III. Grant-free Strategies Enabling Smooth Integration to the 3GPP Standards

In OFDM based systems such as LTE, if timing offsets between UEs are not within the CP length, the signals of the multiple UEs overlap, which increases the eNB detection complexity due to inter-symbol interference (ISI). In LTE, Msg-1 and Msg-2 of the RACH procedure (as illustrated in Fig. 1) would assist the eNB to determine the timing advance (TA) values to synchronize UE transmissions. Since this part of the RACH procedure is omitted in grant-free operation, alternative methods to synchronize UE transmissions in the uplink are required.

We first present the options for uplink synchronization in the absence of the RACH procedure, followed by description of the proposed "grant-free state" for UE operation and its interaction with the legacy LTE states. Then, we elaborate on the NOMA techniques that can be incorporated to smoothly integrate grant-free operation to the 5G NR standards.

A. Uplink Synchronization Options

1) RACH-less long CP (Synchronous): The need for the TA information used to synchronize uplink UEs can be skipped by designing the frame structure using a larger CP length. If the CP is long enough to accommodate the maximum round-trip delay in the cell, it can also accommodate the timing misalignment of the UEs, and still maintain the orthogonality between them. Although having longer CP reduces the actual symbol time and hence the data rate, it is still a feasible solution considering that the IoT devices do not demand high data rates.

2) RACH-less Preamble (Asynchronous): The motivation for asynchronous uplink transmission is a) to further accelerate the uplink transmission speed by transmitting packets in the first step of the RACH procedure (in Msg-1), and b) to further reduce the required signaling messages (eliminate the need for Msg-3, Msg-4, etc.). In the uplink grant-free access, where the TA information is not available, different UE signals will obviously be received at the eNB with different timing offsets, which makes the detection problem more complex. To this end, message-based grant-free access, where the RACH preamble and data are jointly transmitted at different time delays, can be considered. With this method, even if different UEs' signals superpose at the eNB, they can be distinguished by detecting the preamble sequences and corresponding time delays. Moreover, the freedom of selecting different time delays for transmission enables multiple UEs to choose the same preamble, providing another dimension for collision resolution.

In the following, we provide more insights to the asynchronous multiple access and introduce our proposal for a new operating state for grant-free access in the LTE framework.

B. Introducing the new "GRANT FREE" State

In order to aid timing offset estimation and user separation at the eNB, asynchronous multiple access requires a new physical layer design, i.e., new non-orthogonal waveforms, preamble and pilot sequences, robust to time and frequency offsets. NOMA schemes supporting asynchronous operation are limited. One such scheme is the resource spread multiple access (RSMA) [4]. The main challenges in these schemes are:

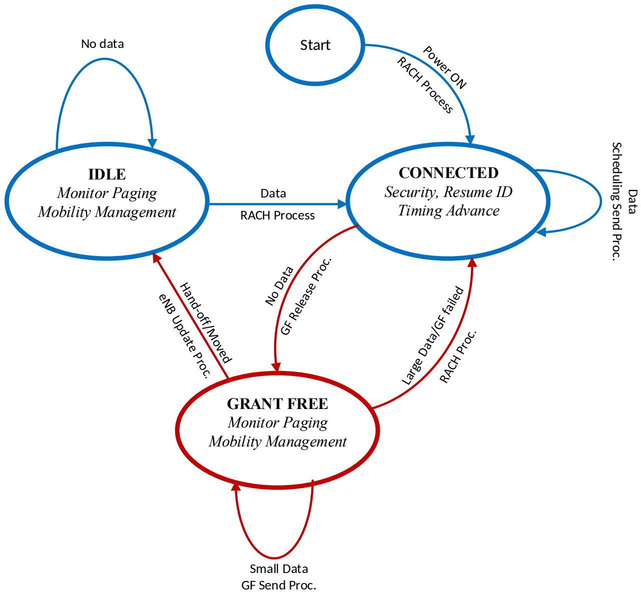


Fig. 2. The interconnection of a "GRANT FREE" state with the "IDLE" state and "CONNECTED" state of a UE. Blue is legacy LTE, red is new "GRANT FREE" state and transitions.

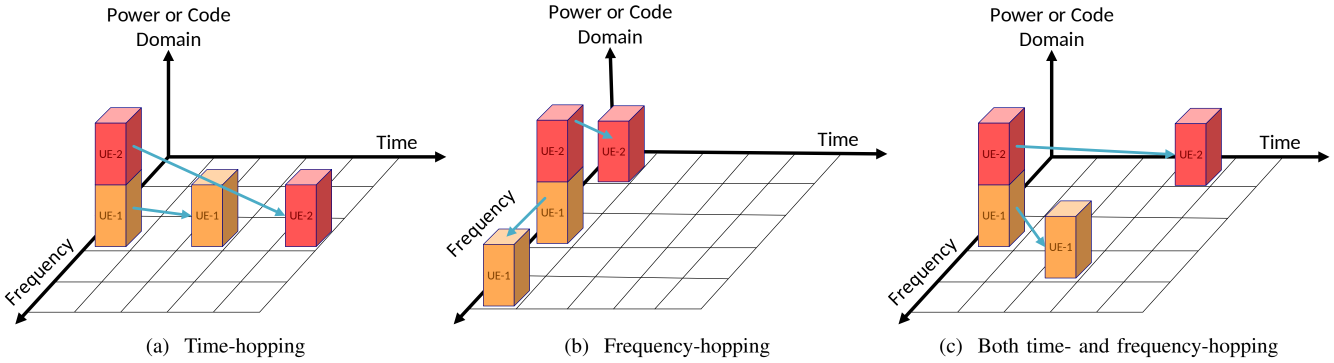
- *Degradation in coverage performance*: Only single tone based RSMA can support the asynchronous case. As with any single-tone method, it suffers from the lack of frequency diversity which can result in bad coverage and increased UE power consumption.
- *Degradation in channel estimation performance*: Channel estimation will be degraded as the uplink pilot sequences or the so-called demodulation reference signals (DMRS) may interfere with data transmissions and orthogonal DMRS multiplexing may not be possible. Additionally, frequency offset correction typically utilizes the DMRS symbols in a subframe and non-orthogonal DMRS transmissions will also degrade the frequency synchronization between the UE transmitter and the eNB receiver.

Therefore, asynchronous multiple access cannot be used at all the stages of UE communication.

To determine the appropriate stage for asynchronous multiple access, we introduce a new "GRANT FREE" state[3]. Fig. 2 shows how the new "GRANT FREE" state fits in with the legacy "IDLE" and "CONNECTED" states. On power up, the RACH procedure is completed to get to the "CONNECTED"

[3] The "GRANT FREE" state is similar to the "INACTIVE" state agreed by the 3GPP RAN2 for LTE standardization. However, the "INACTIVE" state does not segregate the operation based on data size. The UE directly transitions from the "INACTIVE" state to the "CONNECTED" state when data becomes available.

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.



(a) Time-hopping (b) Frequency-hopping (c) Both time- and frequency-hopping

Fig. 3. Back-off options

state and a valid TA is obtained. After completing the data transmission, the UE moves from the "CONNECTED" state to the new "GRANT FREE" state using a grant-free (GF) release procedure. In the "GRANT FREE" state, the UE monitors the paging channel and conducts the mobility management procedures, for example looking for the best cell for potential future communication, which similar to the "IDLE" state. However, unlike the exit procedure from "IDLE" mode, which always requires a high latency, power consuming RACH procedure, the exit procedure from our new "GRANT FREE" state depends on the data size and UE mobility.

For large data transmission, the UE would adopt the legacy RACH-based procedure. But when the amount of data to be sent is small (e.g. $<$ 100 bytes)and if the UE is registered with the same cell, it would adopt an asynchronous NOMA transmission, quickly complete data transmission and return back to the "GRANT FREE" state. This procedure is denoted as "GF Send Procedure" in Fig. 2. If the UE has determined that it has moved away from an eNB and that the TA associated therewith is likely to be inaccurate, it is configured to move from the "GRANT FREE" state to the "IDLE" state. Then, it resumes the legacy RACH procedure on next wake-up to update the TA. However, a UE may not always be able to accurately determine if it has moved away from the eNB. In this instance, the UE executes the GF Send Procedure with an invalid TA, which will likely fail. Then the UE can be configured to directly transfer from the "GRANT FREE" state to the "CONNECTED" state using the legacy RACH procedure upon a predetermined number of failed GF Send Procedures, thereby ensuring a smooth transition from asynchronous to synchronous mode.

If the UE detects a stronger neighboring eNB existence while the UE is in "GRANT FREE" state, it is understood that the TA to the new eNB is unknown. In this case, the UE initiates the eNB Update Procedure. In particular, UE sends a preamble to the new eNB, which responds with a TA and an uplink grant. The eNB may further include operational instructions to the UE, such as to go to the "IDLE" state.

Our concept of a "GRANT FREE" state and the associated transitions are compatible with different candidate NOMA schemes proposed in literature, such as, sparse code multiple access (SCMA), pattern division multiple access (PDMA) and multiuser shared access (MUSA) [2]. However, the use of NOMA in a grant-free environment presents a new set of challenges to be addressed in terms of reliable data decoding. Grant-free NOMA requires advanced schemes for reliable data decoding at the eNB, since the sharing of the time-frequency resources between different UEs results in a higher probability of collisions and interference. A major data decoding procedure considered in this regard is the HARQ, which guarantees reliable uplink transmission by effectively combining the information from initial- and re-transmissions. Next, we summarize the developments in HARQ processing associated with grant-free NOMA, which can be potentially adopted in 5G NR.

C. Enhanced HARQ Capabilities for Grant-free NOMA

The HARQ process for NOMA is more challenging than that in an OFDMA system because these schemes cannot effectively deal with the multi-user interference (MUI), which is prominent in NOMA transmissions. Moreover, due to limited signaling in grant-free transmissions and autonomous transmission of UEs on random resources, the eNB may not know in advance whether a transmission is a new transmission or a retransmission. Furthermore, it may not be able to separate and acknowledge the individual UE transmissions [8]. Therefore, the design of HARQ in uplink grant-free NOMA needs to incorporate novel mechanisms to effectively identify the initial- and re-transmissions and efficiently indicate acknowledgement (ACK) / no-acknowledgement (NACK) messages to the individual UEs. Below, we outline the approaches discussed in the 3GPP standardization committees to resolve these issues [8].

1) Identification and Combining: A simple technique to differentiate initial- and re-transmissions is based on the division of the multiple access (MA) resources into groups. The number of groups would be equal the maximum number of retransmissions [8]. Then, a unique mapping between the resources in each group would be used to determine the transmission type or more specifically the redundancy version (RV) of each transmission. Alternatively, this mapping can also be developed using pilot sequences.

Although this grouping may help to reduce base-station receiver complexity and provide system flexibility, it may

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

inherently reduce statistical multiplexing gains given that small amount of resource would be expected for each group. Hence, it would be desirable to configure a limited number of groups within NOMA resource pool to realize the statistical multiplexing gains while maintaining system flexibility.

2) ACK/NACK: As mentioned earlier, the ACK/NACK mechanism is governed by the efficiency of UE identification and separation at the eNB, which is a challenging task in grant-free NOMA systems. For the UE identification in grant-free NOMA systems, one way is to implicitly transmit the UE identity (ID) piggybacked with the data. Other methods would be to either use a one-to-one mapping between the UE ID and its MA signature or overlay the UE ID on orthogonal DMRS [9]. Regardless of the adopted method, there are three possible cases (and eNB responses) based on detection.

1) Both UE ID and data are successfully detected (eNB sends ACK).
2) eNB failed to detect both the UE ID and data (eNB does not send ACK/NACK).
3) UE ID is successfully detected but the data is not successfully decoded (eNB sends NACK).

For Case 1, the eNB transmits an ACK feedback to the UE. For Case 2 and Case 3, a retransmission is required by the UE. Considering that the UE is in the “GRANT FREE” state (see Fig. 2), it can adopt a grant-based or grant-free retransmission based on the number of failed attempts. However, repeated collisions occur if different UEs keep choosing the same time and frequency resources for retransmission. To eliminate this, the UEs would retransmit in randomly chosen different time and frequency resources using a back-off algorithm. This process is referred to as the asynchronous HARQ. The possible back-off schemes, where the colliding UEs can be separated either in time, frequency or both are shown in Fig. 3.

In summary, all the aforementioned HARQ procedures not only assist in the provision of robust grant-free NOMA solutions, but also have minimal impact on the parent LTE frame structure, ensuring straightforward adoption to the 3GPP standardization for 5G NR.

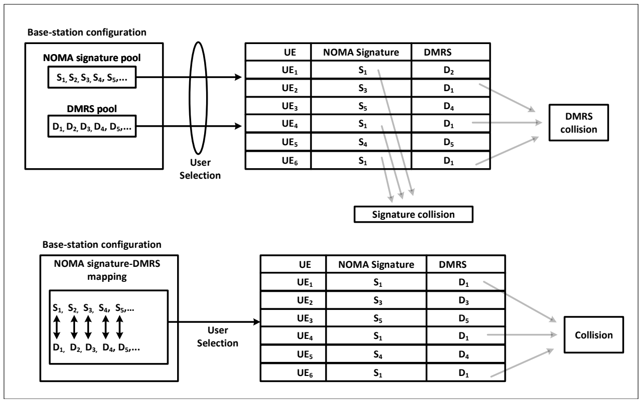
D. Resource Collision Handling for Grant-free NOMA

The two types of grant-free resource allocation methods discussed in 3GPP are:

- Preconfigured/Preassigned Resources: Multiple access (MA) resources consisting of NOMA signatures (e.g., sequence, interleaver, codeword, etc.) and DMRSs are pre-assigned/pre-configured to the UEs via RRC or layer-1 signaling,
- Random selection of resources: When a wireless device has a data to transmit, it randomly chooses a MA resource from a pool of MA resources.

In both types of GF transmissions, MA resource collision occurs when two or more UEs share the same resources. With NOMA in grant-free scenarios, resource collision can occur in the following ways.

- DMRS collision: In this collision, the base station cannot recover the correct channel knowledge resulting in definite decoding failure.


Base-station configuration
NOMA signature pool
$S_1, S_2, S_3, S_4, S_5, \ldots$
DMRS pool
$D_1, D_2, D_3, D_4, D_5, \ldots$
User Selection
DMRS collision
Signature collision
Base-station configuration
NOMA signature-DMRS mapping
$S_1, S_2, S_3, S_4, S_5, \ldots$
$D_1, D_2, D_3, D_4, D_5, \ldots$
User Selection
Collision

UE	NOMA Signature	DMRS
UE_1	S_1	D_2
UE_2	S_3	D_1
UE_3	S_5	D_4
UE_4	S_1	D_1
UE_5	S_4	D_5
UE_6	S_1	D_1

UE	NOMA Signature	DMRS
UE_1	S_1	D_1
UE_2	S_3	D_3
UE_3	S_5	D_5
UE_4	S_1	D_1
UE_5	S_4	D_4
UE_6	S_1	D_1

Fig. 4. Joint vs. separate DMRS and NOMA signature selection.

- NOMA signature collision: In this collision, the data decoding may be successful based on careful signature design and advanced multi-user receivers. However, the collision may still result in decoding failure.

We discuss two solutions being proposed in 3GPP to resolve collision.

1) DMRS Extension: When NOMA is enabled for contention-based grant-free transmissions, the base-station may still perform the UE activity detection and channel estimation based on DMRS. The base-station can pre-filter a short list of active UEs based on the DMRS and apply channel estimation on the short listed UEs. However, in 5G mMTC, the number of UEs may be larger than the maximum number of supported DMRS ports (e.g., 8 or 12 in Release-15). Therefore, one solution is to extend the current DMRS design to support more distinguishable ports by orthogonal extension or non-orthogonal extension. The orthogonal extension can support more distinguishable DMRS ports but at the cost of higher DMRS overhead. In non-orthogonal extension, more distinguishable ports can be generated by using different Pseudo-Noise (PN) sequences or different cyclic shift values. However, the non-orthogonal extension would introduce inter-user interference, impacting the performance of both UE activity detection and channel estimation. Both the topics are currently under research [10].

2) DMRS and NOMA signature selection: DMRS extension is more effective for grant-free transmissions with preassigned resources. For transmissions on randomly chosen resources, procedures to jointly select DMRS and NOMA signatures are being explored. Fig. 4 shows the impact of joint and separate selection of these two entities on the overall collision probability. Separate selection leads to scenarios where signature and DMRS collisions happen on different transmissions, which would increase the overall collision probability. But this problem is eliminated with joint selection and the overall collision probability decreases accordingly [11]. Joint detection has an improved decoding performance [9], reduced collision probability [11], and reduced receiver (e.g., MUD) complexity [12].

A variant of separate selection has been discussed for uplink grant-free SCMA in [13], where the NOMA signatures corre-

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

spond to codebooks and collisions among the same signatures can be resolved by the message passing algorithm (MPA) as long as the channels encountered by the UEs transmitting the same codebooks are different. In a contention based grant-free system with low latency traffic, it is shown that SCMA can provide around 2.8 times gain over OFDMA in terms of supported active users.

If an uplink synchronization is achieved via the methods discussed in Section III.A, similar to the underlying assumption used in the study of SCMA in [13], the base station has to perform blind activity/user detection, resulting in an MUD problem. Our previous work in [14] based on alternative direction method of multipliers (ADMM), demonstrates that the MUD performance can be enhanced by incorporating the signal value estimate from the previous time interval along with the partial active user set as prior knowledge.

Moreover, in [15], the authors have compared the grant-free NOMA and OFDMA under realistic scenarios. As shown in [15], many grant-free schemes, such as SCMA, RSMA and MUSA outperform OFDMA and are more robust to grant-free transmission.

IV. Conclusion

This article provides a detailed description of the standardization aspects related to grant-free transmission and its combination with the 5G enabling technology NOMA. We have discussed the quantification of the key performance indicators being considered by the 3GPP for enhancing the IoT support in 5G mMTC. We have proposed a new state for the grant-free mode of operation and described its coexistence with the legacy LTE states. We also have presented capable grant-free NOMA solutions with respect to uplink synchronization, HARQ procedures and resource collision handling, considering the ease of integration into the current LTE framework. These solutions provide a great advantage for a large number of small packet transmissions, as they can greatly reduce the control signaling overhead, potentially lower the access latency and allow more power efficient operation for low cost devices. In conclusion, standardizing grant-free NOMA mechanisms in 5G NR would provide the much needed substrate for the 3GPP to efficiently host the IoT.

References

[1] J. Zhang, et al., "PoC of SCMA-based uplink grant-free transmission in UCNC for 5G," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1353-1362, Jun. 2017.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. 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.

[3] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, 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.

[4] Y. Wang, B. Ren, S. Sun, S. Kang and X. Yue, "Analysis of non-orthogonal multiple access for 5G," *China Communications*, vol. 13, no. Supplement2, pp. 52-66, 2016.

[5] D. Lin, G. Charbit, and I.-K. Fu, "Uplink contention based multiple access for 5G cellular IoT," *IEEE 82nd Vehicular Technology Conference (VTC Fall)*, pp. 1-5, Sept. 2015.

[6] 3GPP TR 36.881, "Study on Latency Reduction Techniques for LTE."

[7] R1-1610121, "Battery life analysis for grant-free transmission," Qualcomm, 3GPP RAN1 86bis, Lisbon, Portugal, Oct. 2016.

[8] R1-1609039, "HARQ operation for grant-free based multiple access," Samsung, 3GPP RAN1 86bis, Lisbon, Portugal, Oct. 2016.

[9] R1-1807209, "Discussion on NOMA procedure," Google Inc, 3GPP RAN1 93, Busan, Korea, May 2018.

[10] R1-1805909, "Discussion on NoMA related procedures," Huawei, HiSilicon, 3GPP RAN1 93, Busan, Korea, May 2018.

[11] R1-1807075, "Considerations on NOMA related procedures," NTT Docomo, Inc., 3GPP RAN1 93, Busan, Korea, May 2018.

[12] R1-1806307, "Discussion on NOMA procedures," CATT, 3GPP RAN1 93, Busan, Korea, May 2018.

[13] 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," *Globecom Workshops (GC Wkshps)*, pp. 900-905, Dec. 2014.

[14] A. C. Cirik, N. M. Balasubramanya, and L. Lampe, "Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 46-49, Feb. 2018.

[15] Z. Wu, K. Lu, C. Jiang and X. Shao, "Comprehensive Study and Comparison on 5G NOMA Schemes," *IEEE Access*, vol. 6, pp. 18511-18519, 2018.

Ali Cagatay Cirik received the B.S. and M.S. degrees in telecommunications and electronics engineering from Sabanci University, Istanbul, Turkey, in 2007 and 2009, respectively, and a Ph.D. degree in electrical engineering from the University of California, Riverside in 2014. He has worked as an industrial postdoctoral researcher at the University of British Columbia, Vancouver, Canada, and Sierra Wireless, Richmond, Canada between November 2015 and October 2017. He is an inventor in over one hundred pending patent applications. He is currently working at Ofinno Technologies, Reston, VA as a Wireless Technology Specialist. His primary research interests are 5G new radio (NR) and MIMO signal processing.

Naveen Mysore Balasubramanya received the M.S. degree in electrical engineering from the University of Colorado, Boulder, USA, in 2010 and the Ph.D. degree in Electrical and Computer Engineering from The University of British Columbia, Vancouver, BC, Canada in 2017. Since September 2018, he is an assistant professor at the Indian Institute of Technology Dharwad, Karnataka, India. In addition to his academic experience, he has six years of industrial R&D experience. His research focus is on next generation communication technologies and energy efficient Internet of Things.

Lutz Lampe received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the University of Erlangen, Germany, in 1998 and 2002, respectively. Since 2003, he has been with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada, where he is a Full Professor. His research interests are broadly in theory and application of wireless, power line, optical wireless, optical fibre and underwater acoustic communications. Dr. Lampe's contributions have been published widely, and he received of several research and best paper awards.

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Gustav Vos received the BEng degree from the University of Victoria, Canada, and the MEng degree from Simon Fraser University, Canada. He is a chief engineer with Sierra Wireless. He serves as the companys 3GPP representative and is actively involved in developing and directing the required changes to the LTE standard and the 5G standard to optimize it for the Internet of Things. He has 25 years of experience in wide-area wireless communications, having started his career working with proprietary protocols before moving to his current role in advocating for standardized cellular protocols.

Steve Bennett received the BSc (Eng.) degree from the University of London, United Kingdom. He works on R&D for the Internet of Things in the CTO Department, Sierra Wireless. He participates in the 3GPP LTE working group RAN2. He has wide experience in wireless system and product design. He holds 31 patents.

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.