Received 21 December 2019; revised 20 January 2020; accepted 22 January 2020.
Date of publication 28 January 2020; date of current version 13 February 2020.

Digital Object Identifier 10.1109/OJCOMS.2020.2969899

A Survey of NOMA: Current Status and Open Research Challenges

BEHROOZ MAKKI [1] (Senior Member, IEEE), KRISHNA CHITTI [2], ALI BEHRAVAN [3], AND MOHAMED-SLIM ALOUINI [4] (Fellow, IEEE)

[1] Ericsson Research, Ericsson, 417 56 Gothenburg, Sweden
[2] Ericsson Research, Ericsson, 223 62 Lund, Sweden
[3] Ericsson Research, Ericsson, 164 40 Kista, Sweden
[4] Department of Computer, Electrical and Mathematical Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

CORRESPONDING AUTHOR: B. MAKKI (e-mail: behrooz.makki@ericsson.com)

ABSTRACT Non-orthogonal multiple access (NOMA) has been considered as a study-item in 3GPP for 5G new radio (NR). However, it was decided not to continue with it as a work-item, and to leave it for possible use in beyond 5G. In this paper, we first review the discussions that ended in such decision. Particularly, we present simulation comparisons between the Welch-bound equality spread multiple access (WSMA)-based NOMA and multi-user multiple-input-multiple-output (MU-MIMO), where the possible gain of WSMA-based NOMA, compared to MU-MIMO, is negligible. Then, we summarize the 3GPP discussions on NOMA, and propose a number of methods to reduce the implementation complexity and delay of both uplink (UL) and downlink (DL) NOMA-based transmission, as different ways to improve its efficiency. Here, particular attention is paid to reducing the receiver complexity, the cost of hybrid automatic repeat request as well as the user pairing complexity. As demonstrated, different smart techniques can be applied to improve the energy efficiency and the end-to-end transmission delay of NOMA-based systems.

INDEX TERMS 3GPP, 5G, HARQ, MU-MIMO, non-orthogonal multiple access (NOMA), receiver design, user pairing, WSMA.

I. INTRODUCTION

THE DESIGN of multiple access schemes is of interest in the cellular systems design. Here, the goal is to provide multiple user equipments (UEs) with radio resources in a spectrum-, cost- and complexity-efficient manner. In 1G-3G, frequency division multiple access (FDMA), TDMA (T: time) and CDMA (C: code) schemes have been introduced, respectively. Then, Long-Term Evolution (LTE) and LTE-Advanced developed orthogonal frequency division multiple access (OFDMA) and single-carrier (SC)-FDMA as orthogonal multiple access (OMA) schemes. Also, 5G new radio (NR) utilizes OFDMA waveform in both uplink (UL) and downlink (DL) transmission. Such orthogonal designs have the benefit that there is no mutual interference among UEs, leading to high system performance with simple receivers.

In the last few years, non-orthogonal multiple access (NOMA) has received considerable attention as a candidate multiple access technique for LTE, 5G and beyond 5G systems. With NOMA, multiple UEs are co-scheduled and share the same radio resources in time, frequency and/or code. Particularly, 3GPP has considered NOMA in different applications. For instance, NOMA has been introduced as an extension of the network-assisted interference cancellation and suppression (NAICS) for inter-cell interference (ICI) mitigation in LTE Release 12 [1] as well as a study-item of LTE Release 13, under the name of DL multi-user superposition transmission (DMST) [2].

Different schemes have been proposed for NOMA including, power domain NOMA [3], SCMA (SC: sparse code) [4], [5], PDMA (PD: pattern division) [6], RSMA (RS: resource spread) [7], multi-user shared access (MUSA) [8], IGMA (IG: interleave-grid) [9], Welch-bound equality spread multiple access (WSMA) [10], [11], IDMA (ID: interleave-division) [12], NCMA (NC: non-orthogonal coded) [13], ACMA (AC: asynchronous coded) [14], low code rate spreading (LCRS) [15], non-orthogonal coded

access (NOCA) [16], low code rate and signature based shared access (LSSA) [17] as well as UGMA (UG: user grouped) [18]. These techniques follow the superposition principle and, along with differences in bit- and symbol-level NOMA implementation, the main difference among them is the UEs' signature design which is based on spreading, coding, scrambling, or interleaving distinctness.

Various fundamental results have been presented to determine the ultimate performance of NOMA in both DL [3], [19], [20], [21], [22] and UL [21], [22], [23], [24], to incorporate the typical data transmission methods such as hybrid automatic repeat request (HARQ) to the cases using NOMA [25], [26], [27], to develop low-complexity UE pairing schemes [28], [29], [30], and to reduce the receiver complexity [4], [31], [32]. As shown in these works, with proper parameter settings, NOMA has the potential to outperform the existing OMA techniques at the cost of receiver, UE pairing and coordination complexity. For these reasons, NOMA has been suggested as a possibility for data transmission in dense networks with a large number of UEs requesting for access such that there are not enough orthogonal resources to serve them in an OMA-based fashion. Particularly, in 2018, 3GPP considered a study-item to evaluate the benefits of NOMA and provide guidelines on whether NR should support (at least) UL NOMA, in addition to the OMA [33], [34]. However, due to the reasons that we explain in the following, it was decided not to continue with NOMA as a work-item, and to leave it for possible use in beyond 5G.

In this paper, we study the performance of NOMA in UL systems (in the meantime, most of the proposed schemes of Section III are applicable/easy-to-extend for DL transmission). The contributions of the paper are threefold:

- We summarize the final conclusions presented in 3GPP Release 15 study-item on NOMA. Particularly, we present the discussions leading to the conclusion of not continuing with NOMA as a work-item. Such conclusions provide guidelines for the researchers on how to improve the practicality of NOMA.
- We present link-level evaluation results to compare the performance of WSMA-based NOMA and multi-user multiple-input-multiple-output (MU-MIMO) in different conditions. Here, the results are presented for the cases with both ideal and non-ideal channel estimation. As we show, the relative performance gain of WSMA-based NOMA compared to MU-MIMO, in terms of block error rate (BLER), is not that large to motivate its implementation complexity.
- We demonstrate different techniques to reduce the implementation complexity of NOMA-based systems. Here, we concentrate on developing low-complexity schemes for UE pairing, receiver design and NOMA-HARQ, where simple methods can be applied to reduce the implementation complexity of NOMA remarkably. These results are interesting for academia because each of the proposed schemes can be extended and studied analytically in a separate technical paper.

There are a number of survey papers on NOMA [35], [36], [37], [38], [39] in which the performance of power-domain NOMA [35], [36], [37], [38], cognitive radio inspired NOMA [36], code-domain NOMA [38] and signature-based NOMA [39] has been reviewed, and different aspects of MIMO transmission [35], [36], [38], user pairing [37] and receiver design [39] in NOMA have been studied. As opposed, in this paper, we mainly concentrate on the 3GPP discussions on NOMA, comparison between MU-MIMO and WSMA-based NOMA as well as introducing methods to reduce the implementation complexity of NOMA.

As we demonstrate, different techniques can be applied to reduce the implementation complexity of NOMA. Moreover, there is a need to improve the spectral efficiency and the practicality of implementation, in order to have NOMA adopted by the industry.

II. PERFORMANCE ANALYSIS

In this section, we first present the principles of WSMA as an attractive spreading-based NOMA technique. Then, we compare the performance of WSMA NOMA with MU-MIMO and summarize the final conclusions presented in 3GPP Release 15 study-item on NOMA.

A. WSMA-BASED NOMA

WSMA is a spreading-based NOMA scheme [40]. Here, the key feature is to use non-orthogonal short spreading sequences with relatively low cross-correlation for distinguishing multiple users, and the spreading sequences are non-sparse. The WSMA spreading sequences are based on the Welch bound [10], [11], the details of which are explained in the following.

Let us consider K UEs and signals of dimension L. The focus here is limited to symbol-level NOMA where each UE is assigned a UE specific vector from a set of pre-designed vectors. These vectors jointly have certain correlation properties. Consider K vectors, $\{\mathbf{s}_k, k = 1, \ldots, K\}$ called signature sequences (SS), such that each $\mathbf{s}_k$ is of the dimension $(L \times 1)$ and $||\mathbf{s}_k||_2 = 1, \forall k$, where $||\mathbf{s}_k||_2 \doteq \sum_{l=1}^{L} s_{k,l}{}^2$. Let $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]$, be the overall $(L \times K)$ signature matrix. The factor $\frac{K}{L}$ is referred to as the overloading factor in the WSMA context. Since one of the objectives of NOMA is to support a higher user density, it is required to have $\frac{K}{L} > 1$. However, beyond a certain value of $\frac{K}{L}$, the system will be interference-limited. Depending on the required correlation properties of $\mathbf{S}$, a certain performance indicator (PI) is chosen and optimized for the generation of $\mathbf{S}$. One such PI is the total squared correlation (TSC) and is given as

$$\text{TSC} = \sum_{i=1}^{K} \sum_{j=1}^{K} |\mathbf{s}_i^{\text{H}} \mathbf{s}_j|^2, \tag{1}$$

where $(\cdot)^{\text{H}}$ denotes the Hermitian operator. This scalar PI is lower-bounded by a value called the Welch bound (WB) and is given as [10], [11]

$$\text{TSC} \geq \frac{K^2}{L}. \tag{2}$$

On obtaining the optimal value of the chosen PI, TSC in this case, the WB is satisfied with equality and the set $\mathbf{S}$ is called a Welch bound equality (WBE) set. The constituent SSs satisfy WB as an ensemble and not individually, so there exists several sets $\mathbf{S}$ with similar correlation properties satisfying the WB for the same optimal PI. Also, it is required to have a low correlation value, given as $\rho_{ij} = |\mathbf{s}_i^H \mathbf{s}_j|$, between the constituent vectors of $\mathbf{S}$. The motivation to have TSC $= \frac{K^2}{L}$ is that, at the equality, several performance metrics in the system, such as sum-capacity and sum-mean square error (MSE), are optimized simultaneously [10], [11]. This makes it an attractive option for multiple access implementation. Such optimization and SS generation are well understood in the context of interference avoidance techniques [41, Ch. 2].

Other PIs that may be considered for the SS generation include the worst-case matrix coherence given as $\mu = \max_{i \neq j} \rho_{ij}$ and the minimum chordal distance d_{cord} (for detailed mathematical definition see [42]). Optimizing each PI separately will result in a set of SSs each with a different set of correlation properties. Each of these sets is a subset of the WBE set. The number of vectors in each set must be decided before the optimization of the respective PIs. As an example, optimizing TSC will result in a WBE set whose constituent vectors may have unequal correlation among them. Similarly, optimizing the worst-case matrix coherence μ will also produce a WBE set but with an additional property that each constituent SS is equally correlated with every other SS in the set. Such a set is known as a Grassmann set or an equiangular set, and the optimization problem is often referred to as line-space packing problem [43].

At times, it may be required to have zero correlation between few vectors of the constituent SSs in the set. In that case, optimizing d_{cord} is an attractive option. The optimization problem is then referred to as sub-space packing problem [42]. Equations (3)-(5) show the correlation properties of $\mathbf{S}$, each generated by optimizing a different PI, w.r.t the element-wise absolute value of the $(K \times K)$ Grammian matrix ($\mathbf{S}^H\mathbf{S}$) when the number of active UEs $K = 4$. This $\mathbf{S}^H\mathbf{S}$ matrix is independent of the dimension L and is given for different PIs as follows[1]

$$|\mathbf{S}^H\mathbf{S}|_{\text{TSC}} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}, \tag{3}$$

$$|\mathbf{S}^H\mathbf{S}|_{\mu} = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \tag{4}$$

1. WSMA is mainly designed for overloaded systems where $K > L$. However, for simplicity, (3)-(5) show the Grammian matices w.r.t the mentioned PIs at 100% overloading where $K = L$. With $K > L$, the WBE sets are mostly obtained numerically, where the SSs converge iteratively as an ensemble [41, Ch. 2], [44, Table 3]. However, for a few constrained WBE sets, an analytic expression for the SS generation exists, e.g., [45, eq. (11)]. Finally, for examples of the matrices with $K > L$ considered in the simulations of Rel-15, see [33, Appendix A.4].

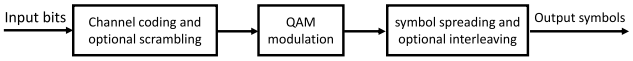


FIGURE 1. Baseband transmitter implementation of WSMA-based NOMA at a user.

$$|\mathbf{S}^H\mathbf{S}|_{\text{d}_{\text{cord}}} = \begin{bmatrix} 1 & 0 & \rho_{13} & \rho_{14} \\ 0 & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & 0 \\ \rho_{14} & \rho_{24} & 0 & 1 \end{bmatrix}. \tag{5}$$

B. NOMA VS MU-MIMO

A generalized block diagram of the baseband transmitter for NOMA implementation is shown in Fig. 1. The information bits of a UE_k are channel coded and then digitally modulated. For a bit-level NOMA implementation, the channel coded bits may be scrambled by a UE specific scrambling sequence and then digitally modulated. Symbol-level UE specific NOMA block appears after the quadrature amplitude modulation (QAM)-modulation block. Using WSMA, each incoming QAM-symbol q_k is repeated L times in a weighted manner by a UE assigned SS $\mathbf{s}_k$ to obtain an $(L \times 1)$ output symbol vector $q_k\mathbf{s}_k$, i.e., a symbol spreading functionality. The repeated symbols $q_k\mathbf{s}_k$ may optionally be interleaved to increase the randomness of the multiuser interference (MUI) to simplify the detector implementation. Usually, $\mathbf{S}$ is pre-generated in the system. To achieve collision-free multiple access, these SSs may be pre-assigned to the UEs, such that no two UEs have the same SS. Cooperation among the UEs may further improve the performance, but comes at an increased complexity and additional communication overhead among the UEs before the actual transmission to the base station (BS).

With K NOMA transmitters, the received $(L \times 1)$ composite vector can be mathematically written as

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{h}_k \odot \mathbf{s}_k \sqrt{p_k} q_k + \mathbf{z}, \tag{6}$$

where for a UE_k, q_k is its QAM-symbol, p_k is its transmit power, and $\mathbf{h}_k$ is the $(L \times 1)$ fading channel to the BS. Also, $\mathbf{z}$ is the $(L \times 1)$ zero mean AWGN vector, and $\odot$ is the element-wise multiplication. Symbol repetition at each UE may be performed in the frequency domain, but it can also be applied in the code domain. With a scalar value p_k, the transmitter allocates the same power to all its incoming QAM symbols. This may be replaced by an $(L \times 1)$ per subcarrier power allocation power vector $\mathbf{p}_k$. Finally, note that (6) assumes that each UE is equipped with a single transmit antenna. However, it may be extended for higher number of transmit antennas, where each spatial layer may have its own SS.

With WSMA, each UE uses L times more resources to transmit the same number of QAM-symbols. This may not be spectrally efficient. This is because without the need to spread the transmit signals, the receiver may at times provide an acceptable performance. This could either be due to the availability of sufficient degrees-of-freedom or due to

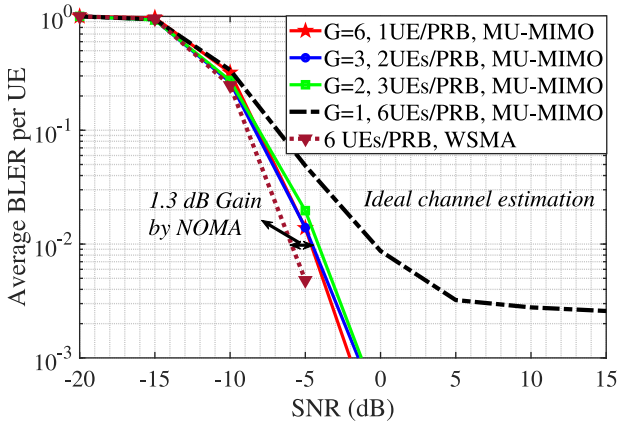


FIGURE 2. NOMA vs MU-MIMO ($K = 6$, an ideal channel estimation).

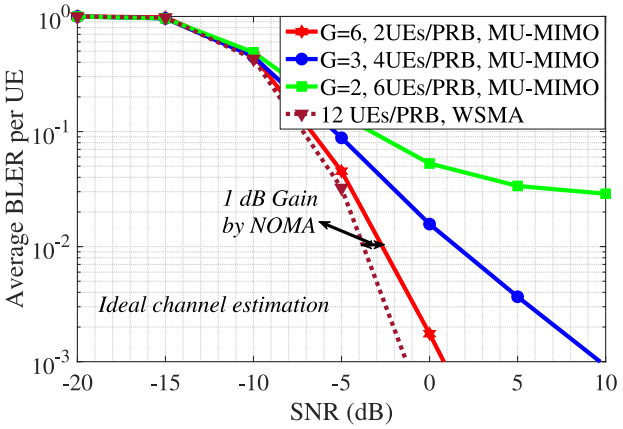


FIGURE 3. NOMA vs MU-MIMO ($K = 12$, an ideal channel estimation).

a good enough interference cancellation at the receiver. In these situations spreading may not be efficient, since it consumes L times more resources, compared to the case with no spreading. Hence it is important to overload the system, i.e., increase K for a fixed L, to increase the sum-rate. This may lead to situations that require optimization of different metrics with conflicting interests.

The baseline system for comparison with NOMA could be an OMA setup when there are $K = L$ users. This is similar to 100% overloading with SS matrix equal to identity matrix of size $K \times K$, and the UEs are scheduled over orthogonal resources. The receiver, BS in this case, receives the composite signal and separates UEs in the frequency domain.

In another case, a baseline system for comparison could be based on MU-MIMO [46], [47]. In this case, both the NOMA and the MU-MIMO systems could be compared for the same number of users per RE. In addition to the frequency domain, the space domain provides additional degrees of freedom (DoF) to the BS. With multiple receive antennas at the BS, a joint space-frequency multiuser detector may be employed. The MU-MIMO system relies only on the spatial separation while NOMA has additional frequency domain, the assumed spreading domain, for UE separation. The same multiuser detector, with a little or no modification in implementation, may be used for both NOMA and MU-MIMO. For the MU-MIMO, an additional UE grouping and scheduling each group over orthogonal REs must also be considered for a fair comparison. Since increasing UE density is one of the NOMA objectives, a comparison for the maximum number of admissible UEs at a given target BLER may also be verified while comparing NOMA and MU-MIMO.

Considering ideal channel estimation, Figs. 2 and 3 show the link-level performance comparison of WSMA with MU-MIMO when the modulation is QPSK and the transport block size (TBS) is 20 bytes. Also, we concentrate on the massive machine type communications (mMTC) scenario of Rel-15 and the inter-BS distance is set to 1732m. The carrier frequency is 700 MHz and we assume that the channels follow the Tapped Delay Line (TDL-C) model [48, Sec. 7.7.2] with the UE moving at 3kmph. Note that TDL-C channel model may be used for simplified evaluation of non-line-of-sight (NLOS) communication. The considered channel model for link-level simulations suffices the NOMA setup which usually targets a high user density coupled with low mobility and small delay spread values. The channel's desired rms delay spread is 30ns. Thus, with the considered speed of the UEs, they experience a flat fading channel. Also, $h_k, \forall k$, are different but the statistics are the same across the UEs. Finally, fading at each UE is independent across slots.

There are four receive antennas at the BS and each UE is equipped with a single antenna. It is assumed that each UE is transmitting with a unit power value over its allocated 6 physical resource blocks (PRBs) and 12 data OFDM symbols. The detector at the BS is MMSE (M: minimum) based. With the spread length 4, the codebook is based on the PI TSC. Here, the spread length refers to the number of resources over which a single transmit symbol is repeated in a weighted manner. That is, for our considered WSMA setup, it refers to L consecutive subcarriers. The channel encoding employed is the rate-matched LDPC code. There is no scrambling and interleaving at the transmitters. AWGN is assumed to have a unit variance. In Figs. 2 and 3, the average BLER values per UE are shown for varying number of UEs $K = 6$ and $K = 12$, respectively. For a given K, to have the same number of UEs per PRB as in the case of WSMA, MU-MIMO divides the UEs into varying number of groups G and varying number of UEs per group N_{u} such that $K = GN_{\mathrm{u}}$.

From Figs. 2 and 3, it can be observed that, for the assumed setup and various values of G, WSMA outperforms MU-MIMO, in terms of BLER, if ideal channel estimation is considered. There is also a saturation observed for MU-MIMO when it is heavily loaded. MU-MIMO systems are interference-limited, i.e., beyond a certain signal-to-noise ratio (SNR), an increase in the transmit power at each user may result in diminishing returns of the performance.

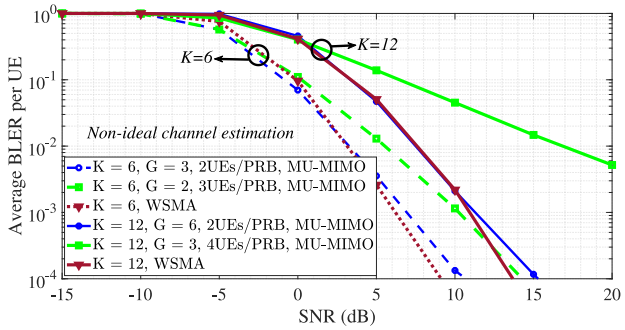


FIGURE 4. NOMA vs MU-MIMO. $K = 6$ and $K = 12$, non-ideal channel estimation.

A user's signal is drowned in the multiple access interference (MAI). With the overloading that NOMA targets, the available spatial DoF in MU-MIMO system are not sufficient to isolate the constituent signals from the received composite signal. This leads to the saturation in the BLER of MU-MIMO.

With NOMA, on the other hand, due to symbol repetition by the low correlation spreading, the energy per resource element (RE) on an average is reduced (note: the SS are unit norm). This ensures that users' signals perceive a lower MAI. This, however, comes at a possible reduced spectral efficiency, since each NOMA UE will consume L times more REs than its MU-MIMO counterpart. This is very prominent at lower overloading factor, where the MU-MIMO outperforms NOMA. Hence a trade-off exists between the overloading and spectral efficiency. Nevertheless, with NOMA the error floors are lowered thereby providing a possibility to squeeze in more UEs per RE for the same target BLER as in MU-MIMO. However, as G increases, MU-MIMO experiences saturation in lower BLERs and the difference between NOMA and MU-MIMO decreases. This is because by increasing G in MU-MIMO, the number of users per RE is reduced, leading to less multiuser interference for each user. More importantly, even with an ideal channel estimation, the relative performance gain of NOMA, compared to MU-MIMO, is negligible, and the relative performance gain decreases with the number of UEs. For instance, considering the parameter setting of Figs. 2 and 3 and BLER 10^{-2}, NOMA-based data transmission reduces the required SNR, compared to MU-MIMO, only by 1.3 and 1 dB in the cases with $K = 6$ and $K = 12$ UEs, respectively. A definite advantage of having a higher UE density with NOMA is visible from Fig. 3.

Figure 4 compares WSMA and MU-MIMO for both $K = 6$ and $K = 12$, but with non-ideal channel estimation. A TDL-A channel model is assumed. At the UEs, the modulation is 16-QAM and the TBS is 60 bytes. As before, the BLER performance of MU-MIMO depends on its configuration, i.e., on the parameters G and N_u. For $K = 12$, the performance of MU-MIMO with $G = 6$ and WSMA is very similar over a wide range of considered SNRs, with the latter outperforming the former only beyond a target BLER of 10^{-3}. When the setup is relatively less dense, i.e., $K = 6$, a similar trend is observed between MU-MIMO with $G = 3$ and WSMA. The target BLER beyond which WSMA performs better is now 10^{-2}. At lower SNR, less than 0 dB, MU-MIMO has a better performance when compared to WSMA. This is also the case for MU-MIMO with $G = 2$. These results do not indicate a possible advantage of MU-MIMO over WSMA and vice versa. However, considering different cases, the performance gain of NOMA, compared to MU-MIMO, is negligible. Finally, while Figs. 2-4 study the average BLER per UE, our simulation results show the same trend when comparing the performance of WSMA and MU-MIMO, in terms of cell throughput [49].

Note that, in the simulations, we have concentrated on WSMA-based NOMA, as an efficient method supporting high user density/throughput with low implementation complexity and little or no modification in the existing OMA-based receivers, e.g., [10], [11], [40]. Then, as shown in [50, Fig. 2], with different receivers and synchronized transmission, the achievable BLER of the WSMA, the MUSA, the SCMA and the RSMA schemes are (almost) the same for a broad range of SNRs. Thus, with fairly high accuracy, the qualitative conclusions of Figs. 2 and 3 hold for these types of NOMA as well. Also, to further improve the receiver's performance, power variation at the UEs could be implemented on top of the spreading based mechanism, e.g., [44], [51]. This will possibly create sufficient variation in the effective channel conditions to enable signal separation. However, for a given NOMA scheme, it has been observed in, e.g., [44, Fig. 15], that the power variation does not offer significant return in gain over the equal power case. Finally, it should be noted that, as opposed to the BSs, in practice the UEs have lower capability/accuracy in power allocation.

C. NOMA FOR BEYOND 5G

During the NOMA Study in 3GPP for 5G NR, a large number of link- and system-level simulations of transmission schemes and corresponding receivers were carried out [33]. In both the link- and the system-level simulations, three scenarios, namely, mMTC, ultra reliable low latency communications (URLLC) and enhanced mobile broadband (eMBB), have been considered with a broad range of parameter settings. Particularly, 14 different companies, each with its own NOMA scheme, provided link-level results and studied the BLER for more than 35 cases. The link-level parameters were generally well aligned among companies, which enabled easy comparison between different methods. Moreover, all NOMA schemes, including those supported by Rel-15, performed similarly at link-level in key conditions. Then, 8 companies provided system-level simulation results, where in total 37 different sets of NOMA versus baseline results were provided. As opposed to the cases with link-level simulations, widely different parameter sets were used in the system-level simulations, with different baselines, making comparisons intractable. Here, the results have been

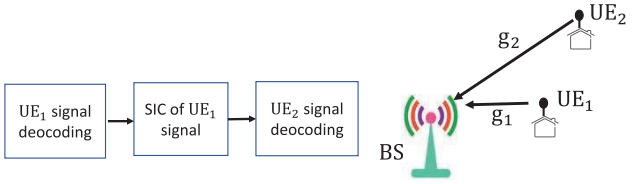


FIGURE 5. UL NOMA. UEs with different channels qualities are paired and the BS performs SIC to decode the signals sequentially.

presented for both synchronous and asynchronous operation models, while the main focus was on the synchronous operation.

According to the results presented during the 3GPP study-item, in ideal conditions, NOMA can be better or worse than MU-MIMO, depending on the number of UEs and simulation parameters. With realistic channel estimation in multipath, however, the relative performance gain of NOMA decreases, and MU-MIMO may outperform NOMA, depending on the parameter settings/channel model.

In a more general point of view, different types of NOMA techniques, including RSMA, IDMA, PDMA, UGMA, MUSA, SCMA, LCRS, WSMA, NOCA, NCMA, IGMA, ACMA, and LSSA, have been considered in the link-level simulations, and their performance have been compared with different existing Rel-15 techniques. However, considering these techniques and different parameter settings/simulation conditions, 1) no specific NOMA technique showed considerably better performance, compared to other schemes, and 2) no clear gain from NOMA over Rel-15 mechanisms were observed in all studied scenarios (see [33, Ch. 8] for details of link-level simulation results). Moreover, in a large number of conditions, the system-level simulations from different companies, mainly concentrating on SCMA, MUSA, RSMA, and PDMA, showed no conclusive gain over Rel-15 techniques (see [33, Ch. 9] for details of system-level simulation results). Also, for all URLLC, eMBB and mMTC scenarios and different NOMA approaches, considerable performance degradation is observed in the cases with non-ideal channel estimation, while the effect of channel estimation is more visible in the URLLC scenario.

In summary, in harmony with our results presented in Figs. 2–4, it was hard to find worthwhile NOMA gains. Moreover, as reported by different companies, the relative performance gain of NOMA, compared to exisiting OMA schemes, is the cost of considerable increment in implementation complexity [33]. These were the main reasons that 3GPP decided not to continue with NOMA as a work-item, and leave it for beyond 5G where new use-cases with ultra-dense UEs may be motivating for NOMA.

III. REDUCING THE IMPLEMENTATION COMPLEXITY

According to the discussions in 3GPP on NOMA, along with the low performance gain of NOMA, compared to existing Rel-15 techniques, one of the key challenges of NOMA is the implementation complexity, in different terms of UE pairing, signal decoding, CSI acquisition, etc. This is specially because NOMA is useful in dense networks where the implementation complexity of the system increases rapidly with the number of UEs. This is the motivation for this section, in which we propose different techniques to reduce the implementation complexity of NOMA. These results are interesting because 1) they provide guidelines to use NOMA with relatively low complexity. Also, 2) each of the proposed schemes, which have been filed in patent applications, can be studied analytically by academia in a separate paper.

For generality and in harmony with the discussions in 3GPP, we present the proposed schemes for UL NOMA. However, as explained in the following, it is straightforward to extend our proposed approaches to the cases with DL transmission. We consider the cases with pairing a cell-center and a cell-edge UE, i.e., UE_1 and UE_2 in Fig. 5, respectively, with $g_1 \geq g_2$ where g_i denotes the channel gain in the UE_i-BS link. However, the discussions hold for arbitrary number of paired UEs. Moreover, for simplicity, we present the setups for the cases with power-domain NOMA and successive interference cancellation (SIC)-based receivers, while the same approaches are applicable for different types of NOMA-based data transmission/receivers. We concentrate on reducing the implementation complexity of the HARQ-based data transmission, UE pairing and receiver as follows.

A. HARQ USING NOMA

Due to the CSI acquisition and UE pairing overhead, NOMA is of most interest in fairly static channels with no frequency hopping where channels remain constant for a number of packet transmissions. As a result, the network suffers from poor diversity. Also, NOMA is faced with error propagation problem where, if the receiver fails to decode a signal, its interference affects the decoding probability of all remaining signals which should be decoded sequentially. For these reasons, there may be a high probability for requiring multiple HARQ retransmissions leading to high end-to-end (E2E) packet transmission delay [25], [26], [27]. The following schemes develop NOMA-HARQ protocols with low implementation complexity.

1) SMART NOMA-HARQ [52]

Our proposed retransmission process is explained in Fig. 6. Assume that in Slot 1 the BS can not decode correctly the signals of the UEs, i.e., $X_1(1)$ and $X_2(1)$. While buffering both failed signals, it asks UE_2 to delay the retransmission of the failed signal. Also, it asks UE_1 (resp. UE_2) to retransmit the failed signal $X_1(1)$ (resp. send a new signal $X_2(2)$) in Slot 2. At the end of Slot 2, the BS first combines the two interference-affected copies of $X_1(1)$ and decodes it using, e.g., maximum ratio combining (MRC). If the signal of UE_1 is correctly decoded, the BS has the chance to use SIC, remove $X_1(1)$ and decode both the failed and the new signals of UE_2, i.e., $X_2(1)$ and $X_2(2)$ received in Slots 1 and 2, respectively, interference-free and with no need for

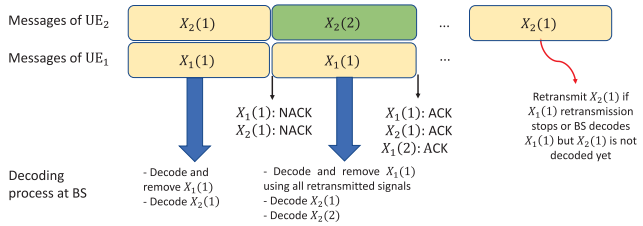


FIGURE 6. Reducing the expected number of retransmissions in NOMA. If the BS fails to decode both signals, it asks for retransmission from only one of the UEs, while the other UE delays the retransmission. The retransmission gives the chance to decode the retransmitted signal. Then, removing the interference, the BS can decode the other failed signal interference-free and with no need for retransmission.

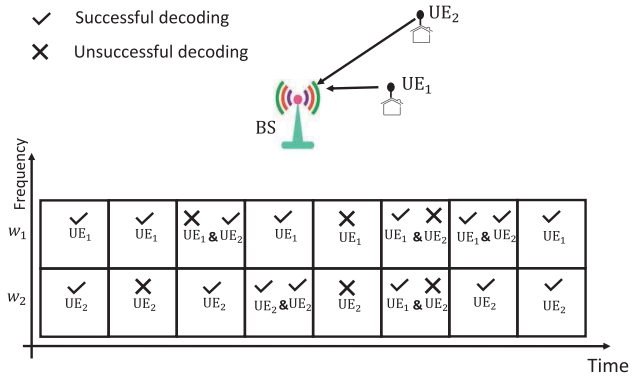


FIGURE 7. Multiple access adaptation in different (re)transmission rounds. If the signal of an UE is not correctly decoded, it has the chance to reuse the spectrum resource of the other UE during retransmissions.

the retransmission from UE_2. Finally, UE_2 starts retransmitting $X_2(1)$ only if the retransmission of UE_1 stops (either because the maximum number of retransmissions is reached or the BS has correctly decoded the signal of UE_1) while the signal of UE_2 has not been decoded yet.

In this way, NOMA gives an opportunity to reduce the number of retransmissions, and improve the E2E throughput. Also, the fairness between the UEs increases because the required number of retransmissions of the cell-edge UE, i.e., UE_2 in Fig. 6, decreases remarkably. The keys to enable such a setup are that 1) the BS should decode all buffered signals in each round and 2) it should inform the UEs about the appropriate retransmission times.

Finally, note that for DL transmission the proposed scheme is adapted as follows. With DL transmission, if none of the UEs can decode their signals correctly, the BS first retransmits the message of the cell-edge UE only, while the cell-center UE receives new messages and buffers the undecoded signals. The cell-edge UE uses typical decoding schemes to decode its own message based on all accumulated signals. On the other hand, following the SIC-based decoding approach, the cell-center UE first decodes and removes the message of the cell-edge UE based on all accumulated retransmitted signals. Then, it tries decoding all of its own received (new and undecoded) signals with no need for retransmissions.

2) DYNAMIC UE PAIRING IN NOMA-HARQ [53]

Here, the objective is to improve the performance gain of NOMA-HARQ by adding *virtual* diversity into the network. In our proposed setup, depending on the message decoding status, different pairs of UEs may be considered for data transmission in different retransmission rounds. As an example, considering Fig. 5, assume that UE_1 and UE_2 with $g_1 \geq g_2$ are paired and send their signals to the BS in a NOMA-based fashion. However, the BS fails to decode $X_1(1)$ (and with high probability $X_2(1)$). Then, in [53], we propose that in the retransmission(s) UE_1 can be paired with a new UE, namely, UE_0 with $g_0 \geq g_1$. This is intuitively because a large portion of the SNR required for successful decoding of $X_1(1)$ has been provided in Round 1. Thus, although we have failed, we are very close to successful decoding and the signal can be correctly decoded by a small boost in the retransmission round. Such a boost can be given by pairing UE_1 with UE_0 having a better channel to the BS. On the other hand, pairing UE_0 and UE_1 gives UE_2 the chance to use a separate resource block to retransmit its own signal interference-free. Also, although the channel coefficients are constant, pairing different UEs in successive rounds provides the BS with different SNR/SINR (I: interference) powers, i.e., diversity, which improves the performance of HARQ protocols considerably. In this way, the fairness between the UEs and the expected E2E packet transmission delay of the network are improved.

Finally, to apply the proposed scheme in DL transmission, the BS can consider a set of predefined UE pairing configurations for each UE. Then, depending on the UEs message decoding conditions, the BS switches to different pairing configurations in different retransmission rounds. Also, with the considered UE pairing and the number of retransmission round, the BS adapts the transmission powers, rates, as well as beamforming and informs the UEs about the considered pairing configuration. The UEs, on the other hand, adapt their decoding scheme based on the instantaneous pairing configuration such that all received copies of each signal are used for message decoding.

3) MULTIPLE ACCESS ADAPTATION IN RETRANSMISSIONS [54]

Our proposed scheme can be well explained in Fig. 7. In our proposed setup, each UE starts data transmission in its own dedicated bandwidth in an OMA-based fashion. Then, if a UE's message is not correctly decoded in a time slot, in the following retransmission rounds it is allowed to reuse the bandwidth of the other UE as well. Let us denote the resource block at time i and bandwidth w_j by $B(i, w_j)$. As an example, consider time Slot 2 in Fig. 7 where the UE_2's message is not correctly decoded while the message of UE_1 is successfully decoded by the BS. Then, in Slot 3, UE_2 uses both w_1 and w_2 to retransmit its message. On the other hand, UE_1 only uses w_1 to send a new message. Using, e.g., repetition time diversity (RTD) HARQ, in $B(3, w_1)$ and $B(3, w_2)$, UE_2 sends the same signal as in $B(2, w_1)$ and the BS decodes the

message based on all three copies of the signal. An example method for message decoding is to first use SIC to decode the message of UE_1 in $B(3, w_1)$, then remove this message from the received signal in $B(3, w_1)$, and use MRC of the three copies of the UE_2's signal to decode its message. Note that, even if the message of UE_1 is not correctly decoded in Slot 3, the BS can still perform, e.g., MRC of the three (two interference-free and one interference-affected) copies of the UE_2's signal.

In this way, compared to the cases with conventional OMA techniques, using the adaptive multiple access scheme, along with HARQ, makes it possible to exploit the network/frequency diversity and increase the UEs' achievable rates. Moreover, our proposed scheme satisfies the tradeoff between the receiver complexity and the network reliability, and, compared to the state-of-the-art OMA-based systems, improves the service availability/the network reliability significantly. Finally, the proposed scheme improves the fairness between UEs and is useful in buffer-limited systems. Also, note that, while we presented the proposed NOMA-HARQ schemes for RTD (Type II) HARQ [55], [56], the same approaches are applicable for other HARQ protocols as well.

B. SIMPLIFYING THE UE PAIRING

With NOMA, optimal UE pairing becomes challenging as the number of UEs increases, because it leads to huge CSI acquisition and feedback overhead as well as running complex optimization algorithms [28], [29], [30]. For these reasons, we present low-complexity UE pairing schemes as follows.

1) RATE-BASED UE PAIRING [57]

Consider a dense network with N UEs and N_c time-frequency chunks where $N_c < N$, i.e., when the number of resources are not enough to serve all UEs in orthogonal resources. An optimal UE pairing algorithm needs to know all N_cN channel coefficients and all N rate demands of the UEs, making the whole system impractical as N and/or N_c increases. This is especially because a large portion of this information is used only for UE pairing and not for data transmission.

To limit the CSI requirement, in [57], we propose that the UE pairing is performed only based on the UEs rate demands and the probability of successful pairing. The proposed scheme is based on the following procedure:

- *Step 1:* The BS asks all UEs to send their rate demands.
- *Step 2:* Receiving the UEs' rate demands and without knowing the instantaneous CSI, the BS finds the probability that two specific UEs can be successfully served through NOMA-based data transmission (see [57] for the detail procedure of finding these probabilities).
- *Step 3:* If the probability of successful pairing for two specific UEs, i.e., the probability that the BS can correctly decode their signals, exceeds some predefined threshold, the BS assigns resources for UL transmission and asks those UEs to send pilots sequentially.

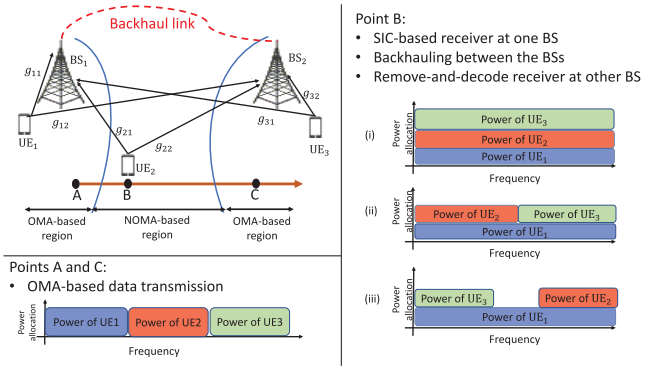


FIGURE 8. UE pairing in CoMP-NOMA. If a cell-edge UE can be successfully paired with a cell-center UE of one BS, it can be paired with each of the cell-center UEs of other BSs with SIC-based receiver only at one BS.

- *Step 4:* Using the received pilots from those paired UEs, the BS estimates the channel qualities in that specific resources, decides if the UEs can be paired and determines the appropriate power level of each UE such that their rate demands can be satisfied.
- *Step 5:* The BS informs the paired UEs about the power levels to use and sends synchronization signals such that their transmit timings are synchronized.

In this way, with our proposed scheme the CSI is acquired only if the BS estimates a high probability for successful UE pairing. This reduces the CSI overhead considerably, particularly in dense networks and/or in the cases with multiple antennas at the UEs. Finally, as we show in [57], to have the maximum number of successful paired UEs, the BS can initially consider the pairs with the highest and lowest rate demands. For instance, assume $r_1 \geq \cdots \geq r_N$ where r_i is the rate demand of UE_i. Then, an appropriate UE pairing approach would be $(r_1, r_N), (r_2, r_{N-1}), \ldots$

Finally, note that the same approach can be applied for DL transmission, except that Step 1 is not required, because the BS already knows the size of the buffered data for each UE. Also, in Step 5, instead of asking the UEs to adapt their transmit powers, the BS informs the UEs to adapt their decoding scheme depending on the considered pairing method.

2) UE PAIRING IN COMP-NOMA

The high-rate reliable backhaul links give the chance to simplify the UE pairing in coordinated multi-point (CoMP) networks using NOMA, e.g., [58]. The idea can be well presented in Fig. 8. Here, depending on the UEs positions and rate demands, they may be served with different multiple access schemes. For instance, in Point A (resp. C) of Fig. 8 where the channel g_{21} (resp. g_{22}) experiences a good quality, UEs transmit in an OMA-based fashion and the BSs may use typical OMA-based receivers with no need for backhauling. In Point B, however, NOMA is used in a CoMP-based fashion and the UEs may share spectrum. As an example, with UE_2 being in Point B, BS_1 may use SIC-based receiver to first decode-and-remove the message of UE_1 and then

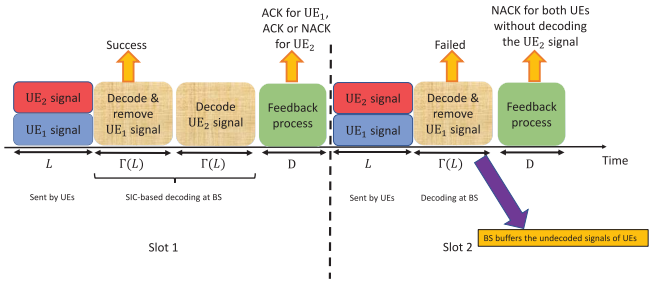


FIGURE 9. Adapting the decoding scheme based on the estimated successful decoding probability. If the message of UE_1 is correctly decoded in Slot 1, the BS continues in the SIC-based receiver scheme to first remove the signal of UE_1 and then decode the message of UE_2 interference-free (see Slot 1). On the other hand, if the BS fails to decode the message of UE_1, it immediately sends NACKs for both UEs without decoding the message of UE_2. Also, the BS buffers the undecoded signals for process in the next rounds of HARQ (see Slot 2). Then, depending on the UEs message decoding status at the BS, in each time slot the UEs' data transmission is synchronized correspondingly.

decode the message of UE_2. Then, using the backhaul link, BS_2 is informed about the message of UE_2 (and, possibly, UE_1) and, removing the interfering signal, it decodes the signal of UE_3 interference-free. Finally, as demonstrated in the figure, depending on the rate requirements and channel conditions, different NOMA-based transmissions with partial spectrum sharing can be considered in Point B, which affect the data transmission, backhauling and message decoding schemes correspondingly.

The advantages of the proposed scheme are: 1) SIC-based receiver is used only in BS_1. Also, 2) UE pairing algorithm can be run only in one of the BSs. That is, NOMA-based data transmission is used as long as at least one of the BSs can find a good pair for UE_2. Finally, 3) pairing (UE_1, UE_2), BS_2 can consider each of its own cell-center UEs to be paired with them as long as the interference to BS_1 is not high. That is, BS_2 does not need to run advanced UE pairing algorithms.

C. RECEIVER ADAPTATION

Compared to OMA-based systems, the sequential decoding process of the BS may lead to large E2E transmission delay, as well as high receiver complexity/energy consumption [4], [31], [32]. Therefore, it is beneficial to use the sequential decoding *only if there is high probability for successful decoding*. This is the motivation for the scheme proposed in the following.

Considering Fig. 9, if the signal of UE_1 is correctly decoded, the BS continues in the typical SIC-based receiver scheme to first remove the signal of UE_1 and then decode the signal of UE_2 interference-free (see Slot 1). On the other hand, if the BS fails to decode the signal of UE_1, it does not continue message decoding and immediately sends NACKs to both UEs without decoding the UE_2's signal. This is motivated by the fact that with NOMA the transmission parameters, e.g., rate, power, of UE_2 are designed based on the assumption that the BS can decode and remove the message of UE_1 and, as a result, it decodes the message of UE_2 interference-free. Then, with an unsuccessful decoding of the UE_1's message, the BS needs to decode the message of UE_2, with poor UE_2-BS link quality, in the presence of UE_1's interfering signal and, with high probability, it fails to decode the message of UE_2 correctly, while it increases the E2E transmission delay. For instance, let us denote the decoding delay for decoding a codeword of length L by $\Gamma(L)$. Then, as shown in Fig. 9, the proposed scheme reduces the E2E delay by $\Gamma(L)$ in Slot 2. The BS buffers the undecoded signal of both UEs for process in the next rounds of HARQ. Then, depending on the decoding approach of the BS and its corresponding decoding delay, in each time slot the UEs' data transmission is synchronized correspondingly.

In this way, the proposed setup reduces the implementation complexity considerably and improves the E2E throughput because the decoding scheme is adapted depending on the estimated probability of successful decoding. Particularly, an interested reader may follow the same method as in [59] to study the E2E performance gain of the proposed scheme analytically. Also, while we presented the setup for the cases with two UEs, it can be shown that the relative performance gain of the proposed scheme increases with the number of paired UEs.

Finally, the proposed approach can be well applied in DL transmission where, if the cell-center UE fails to decode a signal, it stops decoding the following signals and informs the BS immediately. Here, it is interesting to note that, as opposed to the UEs, energy consumption at the BS may not be a problem. Therefore, with an UL transmission the main gain of the proposed scheme is in E2E transmission delay reduction, while it is useful in improving the energy efficiency of the cell-center UE during DL transmission.

IV. CONCLUSION

In this paper, we studied the challenges and advantages of NOMA as a candidate technology in dense networks. As we showed through simulations and in harmony with the discussions in the 3GPP Release 15 study-item on NOMA, NOMA may or may not outperform the typical OMA-based schemes such as MU-MIMO, in terms of BLER. However, for the current use-case scenarios of interest, the relative performance gain of NOMA was not so much such that it could not convince the 3GPP to continue with it as a work-item. On the other hand, the unique properties of NOMA give the chance to develop different techniques reducing its implementation complexity, which may make it more suitable for practical implementation. Therefore, there is a need to improve the spectral efficiency and the practicality of implementation, in order to have NOMA adopted by the industry.

REFERENCES

[1] "Study on network-assisted interference cancellation and suppression (NAICS) for LTE v.12.0.1," 3GPP, Sophia Antipolis, France, Rep. TR 36.866, Mar. 2014.
[2] "New SI proposal: Study on downlink multiuser superposition transmission for LTE," MediaTek, Hsinchu, Taiwan, 3GPP Rep. RP-150496, Mar. 2015.

[3] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "NOMA: An information theoretic perspective," May 2015. [Online]. Available: https://arxiv.org/abs/1504.07751.

[4] F. Wei and W. Chen, "Low complexity iterative receiver design for sparse code multiple access," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 621–634, Feb. 2017.

[5] Y. Yuan and C. Yan, "NOMA study in 3GPP for 5G," in *Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, Hong Kong, Dec. 2018, pp. 1–5.

[6] 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.

[7] *Resource Spread Multiple Access*, 3GPP document R1-164688, Qualcomm, San Diego, CA, USA, May 2016.

[8] 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 Spring)*, Nanjing, China, May 2016, pp. 1–5.

[9] *Non-Orthogonal Multiple Access Candidate for NR*, 3GPP document R1-163992, Samsung, Seoul, South Korea, May 2016.

[10] *NOMA Design Principles and Performance*, document IMT-2020, Ericsson, Beijing, China, Jun. 2017.

[11] X. Li, H.-H. Chen, Y. Qian, B. Rong, and M. R. Soleymani, "Welch bound analysis on generic code division multiple access codes with interference free windows," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1603–1607, Apr. 2009.

[12] C. Xu, Y. Hu, C. Liang, J. Ma, and L. Ping, "Massive MIMO, non-orthogonal multiple access and interleave division multiple access," *IEEE Access*, vol. 5, pp. 14728–14748, 2017.

[13] *Transmitter Side Signal Processing Schemes for NCMA*, document 3GPP R1-1808499, RAN1#94, LG Electron., Seoul, South Korea, Aug. 2018.

[14] *NR-NOMA: Partially Asynchronous and Multi-Layered Transmission of ACMA*, document R1-1808205, 3GPP, Sophia Antipolis, France, Aug. 2018.

[15] *NOMA Transmission Scheme*, document 3GPP R1-1808677, RAN1#94, Intel Corporat., Santa Clara, CA, USA, Aug. 2018.

[16] *System Level Evaluations for NOCA*, document 3GPP R1-1813159, RAN1#95, Nokia, Espoo, Finland, Nov. 2018.

[17] *Signature Generation and Structure of LSSA*, document 3GPP R1-1802069, RAN1#92, ETRI, Daejeon, South Korea, Mar. 2018.

[18] *Transmitter Design of UGMA*, 3GPP document R1-1807073, NTT DOCOMO, Inc., Tokyo, Japan, Aug. 2018.

[19] F. Mokhtari *et al.*, "Download elastic traffic rate optimization via NOMA protocols," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 713–727, Jan. 2019.

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

[21] 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.

[22] 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.

[23] 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.

[24] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using Poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[25] J. Choi, "On HARQ-IR for downlink NOMA systems," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3576–3584, Aug. 2016.

[26] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the performance of NOMA with hybrid ARQ," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10033–10038, Oct. 2018.

[27] Z. Shi, S. Ma, H. ElSawy, G. Yang, and M. S. Alouini, "Cooperative HARQ-assisted NOMA scheme in large-scale D2D networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4286–4302, Sep. 2018.

[28] 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, 2016.

[29] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319–5332, Dec. 2017.

[30] H. Zhang, D.-K. Zhang, W.-X. Meng, and C. Li, "User pairing algorithm with SIC in non-orthogonal multiple access system," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

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

[32] L. Liu, Y. Chi, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving MIMO-NOMA: Iterative LMMSE detection," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1758–1773, Apr. 2019.

[33] "Study on non-orthogonal multiple access (NOMA) for NR," 3GPP, Sophia Antipolis, France, Rep. TR 38.812, Dec. 2018.

[34] *Final Minutes Report RAN185-V100*, document R1-166056, 3GPP, Sophia Antipolis, France, Aug. 2016. [Online]. Available: http://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_86/Docs/

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

[36] 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.

[37] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[38] 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.

[39] 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.

[40] *Signature Design for NoMA*, document 3GPP R1-1802767, RAN1-92, Ericsson, Athens, Greece, Feb. 2018.

[41] D. Popescu and C. Rose, *Interference Avoidance Methods for Wireless Systems*. New York, NY, USA: Springer, 2006.

[42] R. Calderbank, A. Thompson, and Y. Xie, "On block coherence of frames," *Appl. Comput. Harmonic Anal.*, vol. 38, no. 1, pp. 50–71, 2015.

[43] J. A. Tropp, "Complex equiangular tight frames," in *Wavelets XI*, vol. 5914. Bellingham, WA, USA: Int. Soc. Opt. Photon., 2005, Art. no. 591401.

[44] *Transmitter Design for Uplink NOMA*, 3GPP document R1-1809148, NTT DOCOMO, Inc., Tokyo, Japan, Aug. 2018.

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

[46] *MU-MIMO Operation in NR With Configured Grant*, document 3GPP R1-1806248, RAN1-93, Ericsson, Busan, South Korea, May 2018.

[47] *Link Level Evaluations for MU-MIMO*, document RAN1-94 3GPP, R1-1808981, Ericsson, Gothenburg, Sweden, Aug. 2018.

[48] "Study on channel model for frequency spectrum above 6 GHz," 3GPP, Sophia Antipolis, France, Rep. TR-38.900, Jun. 2017.

[49] *On the NOMA Study Outcome*, 3GPP document RP-182577, Ericsson, Stockholm, Sweden, Mar. 2018.

[50] *Link and System Level Performance Evaluation for NOMA*, 3GPP document R1-1802859, Qualcomm, San Diego, CA, USA, Feb. 2018.

[51] *NOMA Scheme With User Grouping*, 3GPP document R1-1802497, NTT DOCOMO, Inc., Tokyo, Japan, Mar. 2018.

[52] B. Makki, M. Hashemi, and A. Behravan, *Smart HARQ in NOMA-Based Networks*, document PCT/EP2018/062393, Ericsson, Stockholm, Sweden, May 2018.

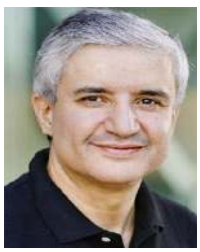
[53] B. Makki, M. Hashemi, and A. Behravan, *Dynamic User Pairing in NOMA-HARQ Networks*, document PCT/EP2018/062456, Ericsson, Stockholm, Sweden, May 2018.

[54] B. Makki, M. Hashemi, and A. Behravan, *Hybrid Automatic Repeat Request Using an Adaptive Multiple Access Scheme*, document PCT/EP2018/053633, Ericsson, Stockholm, Sweden, Feb. 2018.
[55] B. Makki and T. Eriksson, "On the performance of MIMO-ARQ systems with channel state information at the receiver," *IEEE Trans. Commun.*, vol. 62, no. 5, pp. 1588–1603, May 2014.
[56] B. Makki, A. G. I. Amat, and T. Eriksson, "Green communication via power-optimized HARQ protocols," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 161–177, Jan. 2014.
[57] B. Makki, M. Hashemi, and A. Behravan, *CSI-Constrained User Pairing in Dense Uplink NOMA Networks*, document PCT/EP2018/053733, Ericsson, Stockholm, Sweden, Feb. 2018.
[58] K. N. Pappi, P. D. Diamantoulakis, and G. K. Karagiannidis, "Distributed uplink-NOMA for cloud radio access networks," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2274–2277, Oct. 2017.
[59] B. Makki, T. Svensson, G. Caire, and M. Zorzi, "Fast HARQ over finite blocklength codes: A technique for low-latency reliable communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 194–209, Jan. 2019.

BEHROOZ MAKKI (Senior Member, IEEE) received the Ph.D. degree in communication engineering from Chalmers University of Technology, Gothenburg, Sweden. From 2013 to 2017, he was a Postdoctoral Researcher with Chalmers University. He currently works as an experienced Researcher with Ericsson Research, Gothenburg, Sweden. He is a recipient of the VR Research Link grant, Sweden, in 2014, the Ericsson's Research grant, Sweden, in 2013–2015, the ICT SEED grant, Sweden, in 2017, as well as the Wallenbergs research grant, Sweden, in 2018. He has coauthored 56 journal papers, 45 conference papers, and 40 patent applications. His current research interests include partial CSI feedback, hybrid automatic repeat request, green communications, millimeter wave communications, free-space optical communication, NOMA, finite block-length analysis, and backhauling. He is a recipient of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS IEEE Best Reviewer Award in 2018. He currently is an Editor of the IEEE WIRELESS COMMUNICATIONS LETTERS, the IEEE COMMUNICATIONS LETTERS, and the *Journal of Communications and Information Networks*. He was a member of European Commission Projects "mm-Wave Based Mobile Radio Access Network for 5G Integrated Communications" and "ARTIST4G" as well as on various national and international research collaborations.

KRISHNA CHITTI received the master's degree from the Indian Institute of Technology Guwahati, India, in 2010, and the Ph.D. degree from the University of Stuttgart, Germany, in 2015. Since 2018, he has been with Ericsson Research, Lund. His research is focused on signal processing for the PHY layer.

ALI BEHRAVAN received the B.Sc. and M.Sc. degrees in electrical engineering from the Ferdowsi University of Mashad, Iran, in 1994 and 1997, respectively, and the Ph.D. degree in electrical engineering from Chalmers University of Technology in 2006. He has been with Ericsson Research, Stockholm, since 2007, where he is currently working on standardization of new radio. His research interests include wireless access air interface design and evaluations, in particular ultra-reliable low latency communication and industrial Internet of Things.



MOHAMED-SLIM ALOUINI (Fellow, IEEE) was born in Tunis, Tunisia. He received the Ph.D. degree in electrical engineering from California Institute of Technology, Pasadena, CA, USA, in 1998. He served as a Faculty Member with the University of Minnesota, Minneapolis, MN, USA, and Texas A&M University at Qatar, Doha, Qatar. In 2009, he joined King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, as a Professor of electrical engineering. His current research interests include the modeling, design, and performance analysis of wireless communication systems.