Enhancements to the 5G-NR 2-Step RACH: an Unsourced Multiple Access Perspective

Patrick Agostini, Jean-Francois Chamberland, Federico Clazzer, Johannes Dommel, Gianluigi Liva, Andrea Munari, Krishna Narayanan, Yury Polyanskiy, Slawomir Stanczak, and Zoran Utkovski

***Abstract*—The ever increasing demand for massive connectivity, combined with the specific characteristics of machine-generated wireless traffic, is pushing the existing solutions for device registration and data transfer beyond their operational limits. In this paper, we target specifically the 3GPP 5G New Radio random access channel, proposing some enhancements inspired by recent research results in the field of unsourced multiple access (UMAC). Aiming to enable massive grant-free connectivity for large machine-type and Internet of things terminals, we focus on modifications of the two-step random access procedures originally introduced in 3GPP Release 16. By means of simulations, we show that remarkable gains in both energy and spectral efficiency can be attained while retaining a strong resemblance to the legacy protocol.**

I. Introduction

The increasing demand for wireless connectivity and the rise of machine-type communication (MTC) are placing significant strain on existing wireless infrastructure in terms of requests and data transfers. Humans are now connecting to cellular systems through multiple devices, making requests more frequently. At the same time, the number of unattended devices relying on wireless infrastructure for information transfer is growing rapidly. Moreover, the nature of device-driven data transfers, as opposed to human-operated equipment, fundamentally differs from the traditional connectivity-based model. MTC traffic tends to be sporadic and fleeting. This changing dynamic is putting undue stress on the 3GPP 5G New Radio (5GNR) physical random access channel (PRACH), as many more devices are simultaneously competing for spectral resources at any given moment while the resources may be used only for a small fraction of time. In fact, the wireless mechanism used for initial device registration is being pushed beyond the operating range it was originally designed for, rendering the system inefficient and in need of reconsideration. Also the data transmission part requires a re-thinking, as the overhead for resource acquisition becomes a larger burden for the delivery of the smaller data units that typical of MTC. In the past, cell densification has been extensively employed to address issues related to a growing user population. However, with the predicted growth in machine-type data transfers and the shifting trends among existing users, cell densification alone is unlikely to resolve this looming challenge. Some estimates suggest that unattended wireless devices could reach densities several orders of magnitude greater than that of humans. This situation calls for a rethinking of both the resource acquisition and data transmission processes within wireless infrastructures, with a focus on accommodating much greater user densities.

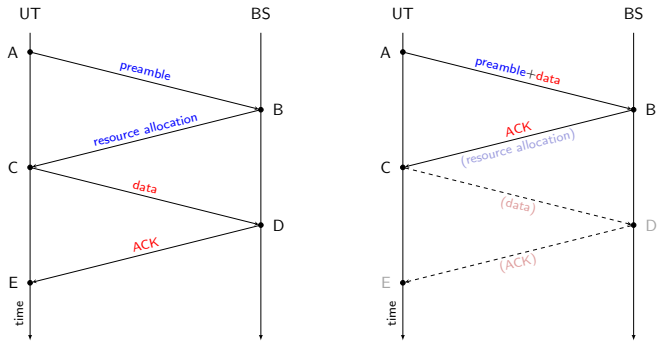
Adopting such mechanisms within next-generation protocols would help prepare wireless systems for the traffic demands of the future. Pragmatic registration and data transmission mechanisms must account for computational complexity and implementability. In this regard, potential registration and data transmission processes are closely related to unsourced multiple access, an area of communication that has garnered considerable attention in recent years. Inspired by these remarks, we propose and discuss in this paper some enhancements to the existing 5G NR two-step random access procedures, showing how remarkable gains in both energy and spectral efficiency can be attained while retaining a strong resemblance to the legacy protocol.

II. 5GNR Two-Step Random Access

We start by reviewing the current registration mechanisms available in 5GNR. The 5GNR random access channel inherits a four-step handshake mechanism from the Long Term Evolution (LTE) standard. From a high-level point of view, the four-step random access (4SRA) protocol works as follows. An active user terminal picks a preamble at random for a set of (up to) 64 Zadoff-Chu sequences. The preamble is sent over the shared physical random access channel (PRACH). At the base station, orthogonal resources—over a physical uplink shared channel (PUSCH)—are granted to each detected user preamble, with the allocation sent back to the user terminal (UT). In the subsequent phase, the UTs transmit their data units over the allocated resources. Upon decoding the transmitted packets, the receiver acknowledges the success to the transmitters. The 4SRA protocol handshake is summarized in Fig. 1a.

With Release 16 of the 5GNR standard [1], a grant-free access mechanism is introduced through the two-step random

P. Agostini, J. Dommel, S. Stanczak and Z. Utkovski are with the Fraunhofer Heinrich-Hertz-Institut, Berlin, Germany. (e-mail: {patrick.agostini, johannes.dommel, slawomir.stanczak, zoran.utkovski}@hhi.fraunhofer.de).
J.-F. Chamberland and K. Narayanan are with the Dep. of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. (e-mail: {chmbrlnd, krn}@tamu.edu).
F. Clazzer, G. Liva and A. Munari are with the Inst. of Communications and Navigation, German Aerospace Center (DLR), Wessling, Germany (e-mail: {federico.clazzer, gianluigi.liva, andrea.munari}@dlr.de),
Y. Polyanskiy is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. (e-mail: yp@mit.edu).
This work was supported by the Federal Ministry of Education and Research of Germany in the programme of "Souverän. Digital. Vernetzt." Joint project 6G-RIC, project identification numbers: 16KISK022, 16KISK056; and, by the National Science Foundation under Grant No CCF-2131115.



(a) Four-step RA (b) Two-step RA

Fig. 1. Random access procedures employed by LTE/5GNR standards. (a) Four-step random access. (b) Two-step random access (Release 16 of the 5GNR standard).

access (2SRA) protocol. Under 2SRA, every active UT picks a random preamble from a set of (up to) 64 Zadoff-Chu sequences, and transmits it over the PRACH. Each preamble points to a resource, in the form of a PUSCH occasion (PO), which is used by the UT to transmit its data unit. The mapping can be one-to-one (OTO) or many-to-one (MTO) [2], [3]. In the OTO case, each preamble points to a distinct PO. In the MTO case, several preambles can point to a same PO, possibly employing different pilot sequences to enable channel estimation even in the presence of collisions. At the base station (BS), the receiver attempts demodulation/decoding at every PO identified by the detected preambles. For decoded packets, the BS acknowledges reception through a feedback message. If a preamble is detected, but the decoding of the associated packet transmission fails, the legacy 4SRA procedure is resumed: the BS signals back an orthogonal resource allocation to the corresponding UT, which proceeds with the retransmission of its packet in the granted resource unit. The procedure of the 2SRA protocol is outlined in Fig. 1b.

A. 5GNR 2SRA and unsourced multiple access (UMAC)

With the perspective of enabling massive grant-free connectivity, we next focus on the first phase of the 2SRA, i.e., we study its performance in isolation, excluding the possibility of exploiting feedback to resume the 4SRA grant-based procedure for unsuccessful packet transmission. This is accomplished by removing ancillary aspects of the protocol such as the specific mapping of the PRACH and POs in the 5GNR framing structure. Specifically, we consider our channel model as a sequence of channel uses which refer to specific time/frequency resources in the 5GNR orthogonal frequency-division modulation (OFDM) grid. Parameter n_{PRE} denotes the preamble size, and n_{PO} is the length of each PO. The number of available POs is given by N. Altogether, the frame length is given by $n = n_{\mathsf{PRE}} + N \cdot n_{\mathsf{PO}}$ real channel uses.

B. Receiver Algorithms

In the remainder of our discussion, we rely on the following detection and decoding algorithms.

- *Preamble detection*: the orthogonal matching pursuit (OMP) algorithm [4], [5] is used to detect preambles. The algorithm is set to provide a list of L preambles. L should be large enough to enable the detection of a moderately large number (in the order of ten or more) of preambles in the early iterations Miss-detections will be filtered out at the decoding stage. For each preamble identified by the OMP algorithm, the steps below are performed.
- *Channel estimation*: A pilot field is included in each PO where a user is attempting transmission and the sequence is determined by the preamble chosen by the user. Denote by $\boldsymbol{x}_{\mathsf{P}}^{(i)}$ the pilot field used by the ith user active in a given PO, and let $\boldsymbol{y}_{\mathsf{P}}$ be the corresponding observation. The estimate of the channel for ith user is obtained as
$$\hat{h}_i = \langle \boldsymbol{y}_{\mathsf{P}}, \boldsymbol{x}_{\mathsf{P}}^{(i)} \rangle / \left\| \boldsymbol{x}_{\mathsf{P}}^{(i)} \right\|_2^2. \tag{1}$$
- *Interference-plus-Noise Power estimation*: Within each PO, the number of active transmissions is unknown to the receiver. Hence, the power to noise plus interference ratio needs to be estimated on a per-PO basis. We use a blind approach and estimate the interference-plus-noise power as
$$\mathsf{NI} = \frac{1}{n_{\mathsf{PO}}} \|\boldsymbol{y}\|_2^2 \tag{2}$$
where $\boldsymbol{y}$ is the observation associated with the PO.
- *Computation of log-likelihood ratios (LLRs)*: LLRs are computed using the channel estimate provided by (1) and the interference-plus-noise power estimate from (2).
- *Decoding*: When packets are encoded with low-density parity-check (LDPC) codes, 50 iterations of the belief propagation (BP) algorithm are employed to decode a message. When packets are encoded with polar codes, successive cancellation list (SCL) decoding is used with adaptive list size (maximum list size 128) [6], [7].
- *Packet Validation*: For LDPC codes, error detection relies on the output of the BP decoder. For the polar code case, a specific error detection method will be discussed in some detail in Section III.

The steps described above, when applied by a receiver, are referred to as treat-interference-as-noise (TIN). We will also consider the improvement obtained by applying successive interference cancellation (SIC) assuming refined channel estimation. When SIC is exploited, a decoded PO and the contribution of the corresponding preamble are removed before the aforementioned steps are repeated.[1]

III. Improving the 5G NR Two-Step Random Access

This section investigates possible modifications to the 5GNR 2SRA protocol aiming at improving its efficiency. We start by studying the potential benefits of using larger preamble sets (Section III-A). Inspired by sparse interleaver division multiple access (IDMA) [8], we then introduce a richer access pattern family that adapts sparse IDMA to the framing structure of 5GNR. We refer to the proposed

[1]Note that the estimation and detection procedures outlined above can be improved, by developing more sophisticated interference-plus-noise power estimators, by performing joint decoding and channel estimation, by adopting more powerful preamble detection techniques, and (in case of non-orthogonal pilot sequences) by replacing (1) with a minimum mean square error (MMSE) channel estimator. This would come at the expense of greater complexity.

scheme as sparse block interleaver division multiple access (SB-IDMA) (Section III-B). Finally, for SB-IDMA, we analyze the performance gain that can be achieved in the short block length regime by replacing the 5GNR LDPC codes [9] with the 5GNR polar codes [10].

A. Extended Preamble Set

A first improvement over the 5GNR 2SRA protocol may be obtained by using preamble sets with cardinality larger than the one of the standardized Zadoff-Chu sequence family. The rationale behind this choice is to remove the tension between the limited multi-packet reception (MPR) capability of OTO configurations (due to the use of identical pilot sequences for colliding users) and the preamble transmission overhead incurred by a repeated use of MTO configurations.

Fig. 3 reports the performance achievable over the quasi-static fading multiple access (MAC) by enlarging the preamble set to 1024 preambles. Therein, we assume each UT transmission is subject to Rayleigh block fading, with fading coefficients that are constant over the whole UMAC frame but independent across UTs. Moreover, non-orthogonal preambles with symbols drawn independently from a complex Gaussian distribution are used. The preamble length is fixed to $n_{\mathsf{PRE}} = 2 \cdot 139$ complex channel uses (c.c.u.s). Each preamble points to a unique pilot sequence of 50 symbols and to one of $N = 64$ POs, realizing a MTO configuration. We consider the (100, 500) 5GNR LDPC codes and quadrature phase shift keying (QPSK) modulation. The rest of the configuration parameters are collected in Table I. Our findings for 1024 preambles show a remarkable gain over the OTO mapping configuration with 64 preambles: the number of active users supported by the system is essentially multiplied by a factor four. The improvement over the MTO configuration with 64 preambles is remarkable, too, although limited to a 50% gain in the number of supported users.

In [11], we tested a setting with up to 16384 preambles and observed signs of saturation at about 100 to 120 supported active users. A possible explanation for this phenomenon lies in the limited MPR capability provided by the error correction scheme and the limited number of access patterns inherent to the slotted Aloha nature of 2SRA. These points are addressed in the following subsection.

B. Sparse-Block Interleaver Division Multiple Access

The modification of the 2SRA proposed in this section is inspired by the sparse IDMA scheme of [12], although with some important distinctions (see [11] for details). The transmission diagram of SB-IDMA is illustrated in Fig. 2. Recalling that the access frame is composed by N POs, the scheme works as follows:

1. The k-bits message $\boldsymbol{u}$ is hashed, generating an index $\phi(\boldsymbol{u})$ that uniquely identifies: (i) a preamble within the preamble dictionary, (ii) a set of n_s pilot sequences, and (iii) a set of n_s distinct indexes in $[1, 2, \ldots, N]$.
2. The message $\boldsymbol{u}$ is encoded via a binary linear block code $\mathcal{C}$. The resulting codeword is modulated using QPSK, yielding n_c-symbol vector $\boldsymbol{c}$.
3. The modulated codeword $\boldsymbol{c}$ is repeated d times.
4. The resulting vector, composed of $d \cdot n_c$ QPSK symbols, is split into n_s segments of equal length.
5. Each of the n_s pilot fields identified in Step 1 is appended to the respective segment, i.e., the first pilot field is appended to the first segment, the second pilot field is appended to the second segment, etc.
6. The preamble selected in Step 1 is appended to the sequence of segments.
7. The preamble is sent over the PRACH. The interleaving pattern selected in Step 1 determines the POs where the individual segments must be transmitted.

The receiver behavior is largely based on the procedures described in Section II-B. Iterative TIN-SIC decoding is assumed. We now discuss some important details. In each iteration of TIN-SIC, observing the PRACH, a set of L preambles is obtained via OMP. For each detected preamble, the sequence of POs used to transmit the user segments is determined, as well as the sequence of pilot fields. In each PO, channel estimation and interference-plus-noise power estimation are performed. LLRs for the codeword bits are computed accordingly. The d LLRs associated with the repetition of each codeword bit are combined (summed) and passed to the decoder of $\mathcal{C}$. Assuming an incomplete decoding algorithm, the outcome of decoding can be either a detected error or a valid codeword decision. In the former case, no further action is needed and the output is simply discarded. In the latter case, an additional error detection step is performed by (i) computing the hash of the decoded message and (ii) comparing it with the preamble index associated with the decoding attempt. If the two indexes are different, the decoder output is discarded. Otherwise, the decoded message is deemed to be correct, and it is passed at the receiver output. For decoded messages that are considered correct, SIC is performed as per discussion in Section II-B.

C. Performance

In the tested SB-IDMA configurations, two types of error correcting codes are considered: the (100, 500) 5GNR LDPC codes and the (100, 500) 5GNR polar codes. We adopt QPSK modulation and introduce 50 symbols per PO as pilots for aiding channel estimation.

In Fig. 3, we consider the quasi-static fading MAC setting. The parameters used for the simulations are given in Table I. The performance of the LDPC-based (—△—) and polar-based (—△—) schemes shows a remarkable gain over the various 2SRA configurations (including the non-standard extension using 1024 preambles). A saturation of the number of supported users, that happens around 160 to 200 users in the polar code case, can still be observed, with a two-fold improvement over the 2SRA OTO mapping with 1024 preambles (- □ -). The saturation appears to be linked to the limited number of preambles, and can be largely improved by enabling an even larger preamble set. However, as observed for the 2SRA case, the number of channel uses allocated to the PRACH (278) limits the number of preambles that can be supported under OMP detection. To enlarge the preamble

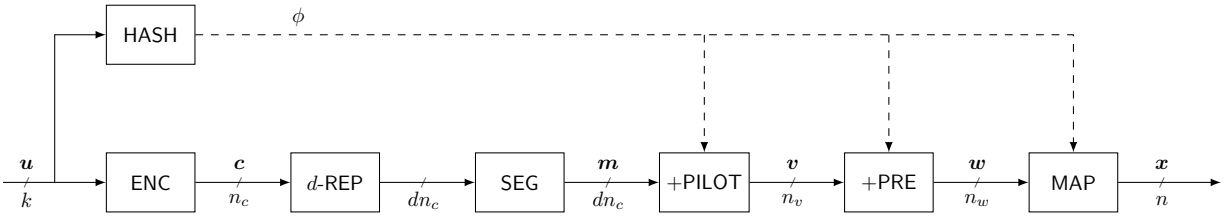


Fig. 2. SB-IDMA transmission chain, as discussed in Sec. III-B.

TABLE I
PARAMETERS USED FOR QUASI-STATIC FADING MAC SIMULATIONS

Parameter	2SRA OTO	2SRA MTO	SB-IDMA LDPC-based	SB-IDMA Polar-based
Frame length (n)	19478	20230	19478	19478
# POs (N)	64	$1 \cdot 35$	64	64
Preamble length	$2 \cdot 139$	$2 \cdot 139$	278	278
# Preambles	64/1024	64	1024	1024
Channel code	LDPC (BG2)	LDPC (BG2)	LDPC (BG2)	Polar (CRC-11)
Repetition rate (d)	-	-	3	3
# Segments (n_s)	-	-	3	3

set, longer preambles can be employed. On the same plot, the performance of an SB-IDMA configuration that trades the number of POs ($N = 59$) for longer preambles (12×139 c.c.u.s) is shown (- -△- -). To limit the energy overhead caused by the longer preamble, a power back-off of 10 dB is applied to the preambles. The resulting performance allows the support of more than 260 active users without signs of saturation.

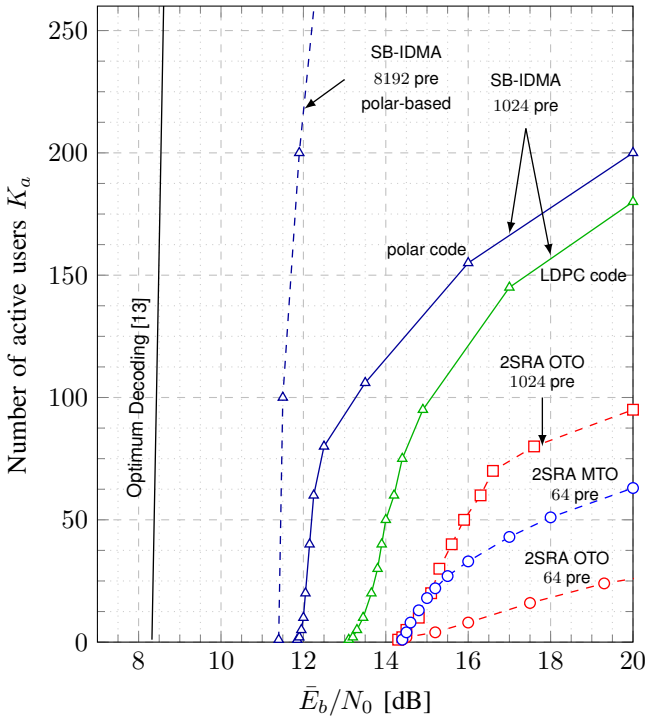


Fig. 3. Number of supported active users vs. average SNR for a target per user probability of error 10^{-1}, as defined in [11]. Quasi-static Rayleigh fading channel, $n \approx 20\,000$ channel uses. Single antenna at the base station.

IV. CONCLUSIONS

This article offers pertinent directions for the random access (RA) to accommodate much larger populations of active users. The proposed schemes leverage notions from IDMA and unsourced random access. Simulation results point to significant performance improvements over existing schemes, thereby enabling wireless systems to operate efficiently under greater user densities. Adopting such access schemes is timely in view of the rising popularity of machine-type communications.

REFERENCES

[1] *5GNR: Medium Access Control (MAC) protocol specification*, 3rd Generation Partnership Project (3GPP) Std. TS 138.321, Rev. 16.1.0, Jul. 2020.

[2] J. Kim, G. Lee, S. Kim, T. Taleb, S. Choi, and S. Bahk, "Two-Step Random Access for 5G System: Latest Trends and Challenges," *IEEE Netw.*, vol. 35, no. 1, pp. 273–279, Jan./Feb. 2021.

[3] E. Peralta, T. Levanen, F. Frederiksen, and M. Valkama, "Two-Step Random Access in 5G New Radio: Channel Structure Design and Performance," in *Proc. Vehicular Technology Conference*, Apr. 2021.

[4] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, US, Nov. 1993.

[5] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[6] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015.

[7] B. Li, H. Shen, and D. Tse, "An Adaptive Successive Cancellation List Decoder for Polar Codes with Cyclic Redundancy Check," *IEEE Commun. Lett.*, vol. 16, no. 12, p. 2044–2047, Dec. 2012.

[8] A. K. Pradhan, V. K. Amalladinne, A. Vem, K. R. Narayanan, and J.-F. Chamberland, "Sparse IDMA: A Joint Graph-Based Coding Scheme for Unsourced Random Access," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7124–7133, Nov. 2022.

[9] T. Richardson and S. Kudekar, "Design of low-density parity check codes for 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 28–34, 2018.

[10] V. Bioglio, C. Condo, and I. Land, "Design of polar codes in 5G New Radio," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 29–40, 2020.

[11] P. Agostini, J.-F. Chamberland, F. Clazzer, J. Dommel, G. Liva, A. Munari, K. Narayanan, Y. Polyanskiy, S. Stanczak, and Z. Utkovski, "Evolution of the 5G New Radio Two-Step Random Access towards 6G Unsourced MAC," 2024. [Online]. Available: https://arxiv.org/abs/2405.03348

[12] V. K. Amalladinne, J.-F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for unsourced multiple access," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6509–6533, Oct. 2020.

[13] S. S. Kowshik, K. Andreev, A. Frolov, and Y. Polyanskiy, "Energy efficient coded random access for the wireless uplink," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4694–4708, Aug. 2020.