Multiplicity Estimating Random Access Protocol for Resource Efficiency in Contention based NOMA

H. Murat Gürsu [ℑ], Berkay Köprü [⋆], Sinem Coleri Ergen [⋆], Wolfgang Kellerer [ℑ]
[ℑ]Chair of Communication Networks, Technical University of Munich, Germany
[⋆]Wireless Network Laboratory, Koc University, Istanbul, Turkey
Email: [ℑ]murat.guersu@tum.de, [⋆]bkopru17@ku.edu.tr

***Abstract*—Emerging technologies enforce strict requirements on future wireless networks such as massive connectivity that cannot be supported with scheduled access. Contention based Non-Orthogonal Multiple Access is a novel technique to overcome strict massive connectivity requirements by efficient use of wireless resources. However, most of the solutions proposed in this direction assumes different loads which would degrade the performance significantly if they would not hold. To stress these assumptions a resource efficiency metric is defined and state of the art solutions are evaluated for varying load regarding this metric. It is shown that the resource efficiency problem in the state of the art can be improved with multiplicity estimation, and hence, we propose Multiplicity estimating Random Access protocol, that adapts to the dynamic loads. This adaptation is evaluated through analytical calculation against the state of the art and it is shown that resource efficiency against with a slight decrease in the metric any load from 1 up to $> 10^3$ users is supported . In addition, we show how this protocol can be dimensioned and integrated to contention based NOMA.**

I. Introduction

Non-Orthogonal Multiple Access (NOMA) is known to achieve higher spectral efficiencies compared to orthogonal multiple access alternatives. However, the price to pay is the computational complexity resulting in higher iterative signal processing requirements at the receiver. The unleash of the higher computational power, via distributed and cloud computing [1], is a possible solution to this problem and NOMA is a candidate technology for the next mobile network technology. Hence, it is a promising solution to the massive access problem of upcoming Machine Type Communication (MTC) [2].

NOMA schemes enable use of same resource for multiple users through different techniques such as successive interference cancellation. Sharing of a resource such that multiple packets can be received on that resource at the same time is refereed as K-Multi Packet Reception (K-MPR) [3]. K-MPR is a superset of multiple access schemes exploiting orthogonal resources which is in the sense 1-MPR as there is a single packet reception per resource. In case the users are not scheduled, they can take distributed decisions to use these resources on the fly. Such distributed decisions can result in more than one packet on a resource, which is referred to as collisions. Due to this limitation random access systems are built on a three state model with idle, singleton and collisions. However, with K-MPR we have $K+2$ states for resources in random access schemes as each packet reception from 1 to K packets is considered a different state. Thus, a larger set of successful states are available compared to 1-MPR. Although the success is possible for a wider range of decisions it is still possible with more than K users accessing the same resource simultaneously such that collisions are not fully avoided. With respect to this reason recent work [4] has suggested NOMA as a solution for contention based access. However, there it is assumed that the users cardinality is limited and a collision never occurs. This cannot be guaranteed unless a backlog estimation is deployed. These concerns were already highlighted in the state of the art [5].

Prior works have dealt with the contention based NOMA problem in three ways. Firstly, the user multiplicity is assumed to be known [6]. This is realizable if the amount of available resources is always larger than the number of users. This constraint, although possible in small cells limits the generalization. There is also other branches of work that infer the user multiplicity through timing differences of each user [7] or that extract certain correlations from traffic characteristics [8]. Second, since the channel estimation improves the NOMA performance, massive number of pilot allocation is used as an indirect way to solve the multiplicity problem [9]. Third, users that have activated the same pilot can be separated using channel correlations [10] or through capture effect [11]. This again is similar to the former assumption but more realistic since pilots are less costly compared to NOMA resources. However, the number of required pilots can increase to the resource limitation of the system where it is not feasible to have non-contaminated pilots anymore [12].

In this paper, we claim that NOMA needs a random access protocol to be adaptive for MTC. This random access protocol is not based on a certain request type but, rather on the multiplicity and the channel estimation of the devices as motivated with earlier work. In order to motivate the traffic limitation of previous contention based NOMA solutions we start by defining a resource effiency metric. Following, we suggest a new random access protocol for NOMA-MTC. The contribution of the paper are twofold:

1) We introduce a resource efficiency metric for contention based NOMA through which we show that state of the

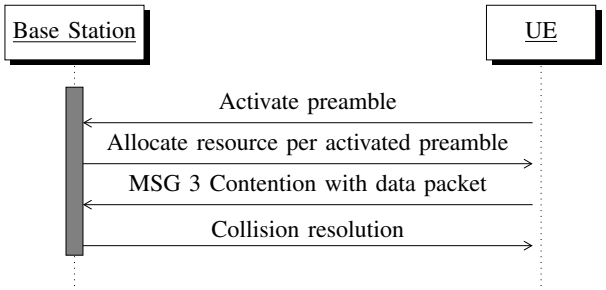

Fig. 1: Message sequence chart for LTE Random Access

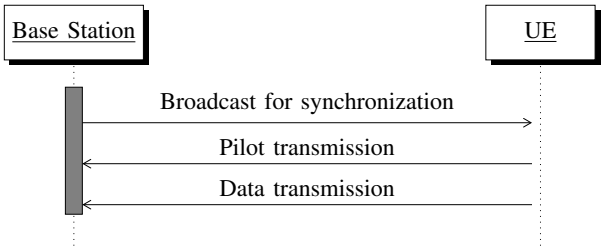

Fig. 2: Message sequence chart for Grant-Free Access for NOMA

art protocols are designed for certain arrivals types.
2) Through the guidance of the metric we show that multiplicity estimation extends the operation region for varying traffic compared to the state of the art.

The paper is organized as follows: In Sec. II we summarize the state of the art and emphasize the contention based NOMA problem. We explain the background and the scenario in Sec. III. Following we propose a protocol that can overcome such a problem in Sec. IV. The Resource Efficiency metric is defined and the proposed protocol is evaluated against the state of the art in V. The paper is concluded in Sec. VI.

II. Related Work

Channel information collection is rarely emphasized in the contention based NOMA schemes. Practical implementations that have limited channel information are investigated in [6]. However, with corrupted information the performance degrades and efficiency of NOMA techniques decreases. In order to deal with this problem there are two different approaches: Contention Resolution and Grant-Free. First the former, then the latter will be explained in the following.

A. Contention Resolution

Contention resolution based random access procedure is already used in the latest generation of the mobile networks Long-Term Evolution (LTE). In LTE random access, when a user needs to access a resource, it transmit a preamble on the pre-allocated Random Access Channel (RACH) resources. This is illustrated in Fig. 1. The base station receives the set of activated preambles and responds with OFDMA resources and transmission power settings for these preambles. The users then transmit on the allocated resources. If multiple users have selected the same preamble simultaneously, a collision occurs and the base station cannot correctly receive the transmission. At this moment a collision resolution algorithm is activated to separate users to different resources. This takes certain time depending on the load of the random access channel [13].

A random access procedure has been proposed for NOMA communication in recent work [7]. In this work a random access procedure is used for obtaining the channel estimation through use of preambles, after which users are allocated resources in NOMA blocks. The channel estimation may fail after the preamble transmission, in which case users randomly back-off and re-transmit new preambles. The study has multiple limitations: First, the traffic is assumed to be known such that the preamble allocation periodicity is adjusted accordingly. Secondly, it is assumed that up to two overlapping preambles can be differentiated due to time or arrival difference. This imposes space distribution or synchronization assumptions on the users. Another work [14] used the same assumption for the channel estimation for the preambles attached to the starting of the packet.

In [9] the focus is on uplink contention for sparse code multiple access (SCMA) as another flavor of NOMA. In this work, the channel estimation is assumed to be resolved via back-off and no further evaluation of this is shared. Contention resolution is one way to deal with contaminated channel estimation resources. Another branch of the work assumes do deal with this problem through over-dimensioning as in Grant Free approach.

B. Grant-Free

In [15] the authors present an emulation of random access based NOMA. The procedure is that users receive a synchronization signal and transmit their pilot and data successively as illustrated in Fig. 2. In their work, the authors have implemented an over-dimensioning of channel estimation resources. These resources are referred as pilots, in an over-dimensioning ratio of 1 to 8 pilots per maximum number of active users at any time instance. Such an over-dimensioning assumes that the maximum number of users accessing the system at a time instance is limited and known. This assumption can be challenged in cells with varying traffic profile. It is also possible to impose further assumptions on traffic to deal with the channel estimation problem. In [8] the authors make use of the temporal correlation of the transmitted packets from the same user to infer the channel estimation. They provide a compressive sensing based technique that helps the successive interference cancellation technique to improve the performance of NOMA. This puts a further assumption on the traffic, which is the point that we want to avoid.

III. Background

In this section we introduce the state of the art for the elements used for the introduced protocol.

A. Multiplicity Estimation

In wireless random access the number of users n, that are simultaneously accessing the resources, is unknown. Usually,

it is assumed that the n is a realization from a well-known probability distribution and this distribution is used to dimension the system. There is also a line of work that treats n as an estimation problem and multiplicity estimators are proposed for different applications, one example is the work in [16] for RFID systems. The main idea is to use an estimation over the probabilistic outcomes of busy and idle resources over l slots. The users activate a subset of l slots and the percentage of the active slots can be used to estimate the number of active users n. Given the number of active users n and the number of resources l are sufficiently large the multiplicity of the users can be detected precisely.

B. Channel Estimation

Previous work [11] have considered a strongest user collision resolution for detection of a single user over used pilots. This is similar to capture effect discussion in [17] which is based on assumptions on path loss and transmission power. In our scenario we consider no assumption on received power diversity. Thus, if two or more users select the same pilot we consider it as a false positive detection of a single user on that pilot and act accordingly.

C. Scenario

We are interested in an uplink scenario in a star topology where many sensors are transmitting their messages to the central entity, dubbed as base station. We consider that there is a resource grid formed of chunks of time and frequency. We call any grouping within the grid as resources and the smallest unit of resources is referred to as symbols. A pilot is a group of multiple symbols that is used for detection of the channel characteristics, the length of a pilot is not fixed and specified for different protocols in the evaluation section. There are in total N users in a single star topology and we are investigating a specific time instance where only a subset of these users are active with cardinality n. We assume that the channels are varying from previous activation to next activation time of the sensor. Thus, it is necessary that the channel is estimated again. Each pilot has a NOMA resource assigned to it and multiple pilots can be assigned to the same NOMA resource to benefit from the K-MPR capabilities.

In this work we analyze the impact of over-dimensioning and the assumptions imposed on the arrivals via formulating a performance metric. Furthermore, we suggest a new random access protocol for NOMA that would deal with varying arrivals in a resource efficient way against a wider range of traffic profiles.

IV. Multiplicity Estimating Random Access Protocol

The proposed Multiplicity Estimating Random Access Protocol (MERAP) protocol has three steps, user multiplicity estimation, channel estimation and resource use. This is illustrated in Fig. 3. When a unknown number of users n have a packet to transmit, they each transmit a single symbol in the multiplicity estimation frame. The multiplicity estimation

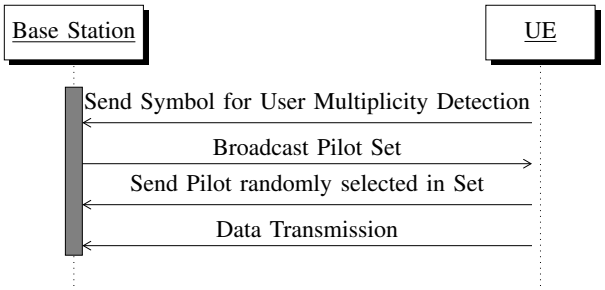

Fig. 3: Message sequence chart for proposed random access protocol for NOMA

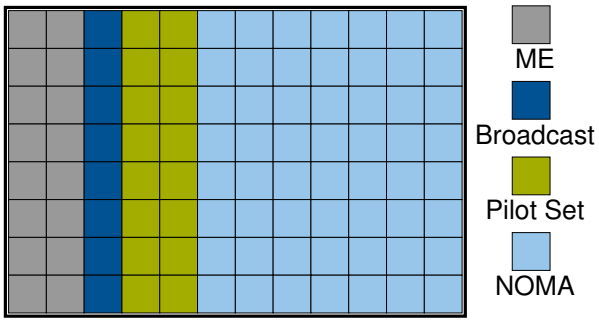

Fig. 4: The resource grid required for the proposed random access protocol for NOMA. The x-axis denotes time chunks while the y-axis depicts subcarriers. We will refer to the repeated pattern of this grid as the NOMA frame.

frame is composed of l symbols from only one is selected by each user. This selection is done randomly in uniform fashion. The base station reads which symbols are transmitted out of l symbols and uses the quantity to calculate the load $\hat{\xi}$. This information is fed to an estimator for the multiplicity estimation.

Following, the base station defines a set of m pilots for all active users and informs users through a broadcast. The users are aware of which m pilots they can select from but they are not allocated a unique pilot. Thus, each user select one of these m pilots randomly in a uniform fashion. After that the user can transmit the data on the related NOMA resource. The relation of which resource is attached to which pilot can be preset. Such a protocol can function in a repeated frame structure. This is illustrated in Fig. 4 where broadcasts are downlink and all the other resources are used in an uplink way. Following, we investigate the Multiplicity Estimation (ME).

For the multiplicity estimation part we assume a binary channel model with states idle and busy. Each symbol can be detected as busy or idle by the base station. We assume that this is enabled via a low enough noise level, no external interference and sufficient transmission power.

A. User Multiplicity Detection

The users can start the Multiplicity Estimation (ME) protocol directly via a single symbol transmission as in LTE. As informed previously about the location of the ME frame all

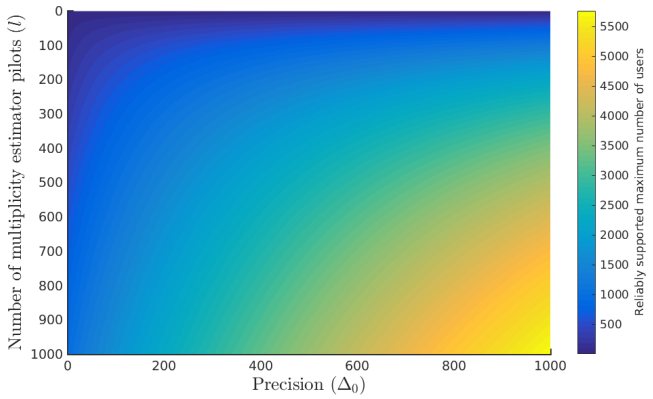

Fig. 5: Reliably supported average users n_{max} against number of multiplicity estimation resources l and the estimation constraint.

active users select uniformly and randomly to transmit one of the l symbols in ME. At each symbol busy or idle state can be detected by the base station. Depending on the reception of the idle and busy states of the symbols on the ME, the base station can count the idle symbols l_I, which can be used to calculate the estimated load $\hat{\xi}$ i.e., number of active users $\hat{n}$ per multiplicity detection resource l in a NOMA frame $\hat{\xi} = \frac{\hat{n}}{l}$ as in,

$$l_I = l \cdot \left(e^{-\hat{\xi}}\right). \tag{1}$$

This can be used to calculate the load in a closed form as

$$\hat{\xi} = \ln\left(\frac{l}{l_I}\right), \text{thus } \hat{n} = l \cdot \ln\left(\frac{l}{l_I}\right) \tag{2}$$

where ln is the natural logarithm. This is the mean estimation of a Poisson distribution from l realizations. We can use a Gaussian distribution for approximating a Poisson distribution with large mean values. Then, we calculate the variance of this distribution. This perspective is introduced in [16] where the standard deviation is given as

$$\sqrt{n \cdot \frac{e^{\xi} - (1+\xi)}{\xi}} = \sigma(n, l). \tag{3}$$

As we are building a distribution for the estimation, the lower the standard deviation, the sharper is our estimation. We can use the confidence interval r_{me} to provide the mean of the distribution as our estimate. We define precision $\Delta(n, l)$ as the 3 sigma limit $\Delta(n, l) \triangleq 3\sigma(n, \xi)$ that results in $r_{me} \approx 0.99$.

For a fixed l and a constant precision Δ_0, we can calculate the maximum number of users n_{max} that can be precisely detected as

$$n_{max} = \Delta_0 + \arg\max_{n} n, \text{ s.t. } \Delta_0 \geq \Delta(n, l). \tag{4}$$

The Δ function is not closed-form invertible. However, it can be evaluated for various n and l values such that for any given threshold Δ_0 and l the supported n_{max} can be tracked back.

p_{ce}	0.9	0.99	0.999	0.9999
m per user	9.4912	99.4992	999.4999	9999.5

TABLE I: The relation between the reliability requirement and over-dimensioning of pilots.

We represent this search from the calculated values with a function $\mathrm{f}(l, \Delta_0)$ and re-write the $n_{max} = \mathrm{f}(l, \Delta_0)$.

The function $\mathrm{f}(.)$ is plotted in Fig. 5 with varying Δ_0 and l. We see that the precision threshold is limiting the maximum number of users the estimator can detect with the given confidence. Clearly, to support more users the precision of the estimate should be relaxed. This results in decreasing multiplicity estimation resources. However, the trade-off is the resulting over-dimensioning for channel estimation resources.

B. Pilot Allocation

After the estimation, the pilot set size m is defined and broadcast to users. Users select one pilot from the allocated set uniformly. If the size of the pilot set is m, using the Poissonization of the binomial distribution, we can write the probability of a pilot being selected by a single user as

$$p_{ce} = e^{-\lambda_{ce}} \tag{5}$$

where $\lambda_{ce} = \frac{n}{m}$ is the user per pilot density. Since we do not have the actual user density we can use the estimated value and variance to allocate the pilot set such that $\lambda_{me} = \frac{\mathrm{f}(l,\Delta_0)}{m}$.As there is no means to detect collisions in pilots the price to pay is the idle slots. It is important to note that we did not use the probability to detect a single user on one pilot. Thus, the size of pilots and the size of the multiplicity detection frame creates a trade-off in terms of system resources for the same purpose which is the successful pilot probability. This trade-off is shown in Tab. I with various constraints. Over-dimensioning with respect to required reliability is increasing from 1 to 10 to 1 to 10000.

We set the limit the channel estimation resources as $p_{ce} > 0.99$. And then we use

$$\arg\min_{m} \quad e^{-\frac{\mathrm{f}(l,\Delta_0)}{m}} > 0.99, \tag{6}$$

to set the optimal required number of resources.

This calculation is done in real-time after every multiplicity estimation to allocate optimal number of pilot resources. Then, this information is broadcast to the users.

C. Resource Use

Each pilot is assigned to a different NOMA resource, thus, a user that have selected a pilot knows on which NOMA resource it should transmit. The only case of problem will be a pilot with multiple users where the channel is estimated erroneously, referred to as false positive. These allocations will result in wasted NOMA resources. Since a resource is available per every actively detected pilot even though it is a false positive we can calculate the amount of resources allocated for NOMA as $G_n = \alpha_d \left(1 - e^{-\lambda_{ce}}\right) m$ where α_d is the number of symbols required per NOMA resource.

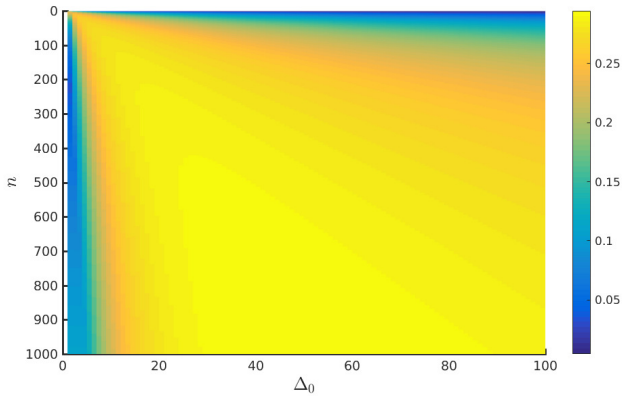

Fig. 6: The efficiency with respect to varying the multiplicity estimation threshold and number of users with $r_{me} = 0.9$.

V. Evaluation

In this section we want to evaluate the resource efficiency of the state of the art solutions and the proposed solution to emphasize the operating regions.

A. Resource Efficiency Metric Definition

The efficiency ρ is defined as the data resources divided by the sum of total resources used i.e., multiplicity estimation resources G_{me}, channel estimation resources G_{ce}, NOMA resources G_n and successful NOMA resources G_{ns} as in

$$\rho = \frac{G_{ns}}{G_{me} + G_{ce} + G_n}. \tag{7}$$

In order to provide comparability we express every resource in terms of symbols.

B. Protocols

We calculate the efficiency of the state of the art through the information provided in the papers mainly in PoC [15] and NORA [7]. The traffic in PoC is Poisson distributed with a mean of 3 arrivals per subframe and Beta distributed as given in [13] for NORA. As the multiplicity estimation is not used the resource used is set to zero $G_{me} = 0$. The parameters used in the comparison is summarized in Tab. II with their explanations.

1) PoC: In PoC 4 symbols are used per subframe to allocate 48 pilot sequences. The size of pilots are denoted as m and the number of active users n. In [15] the collision probability p_c of u^{th} user is given as

$$p_c(u,m) = \begin{cases} 1 - \prod_{i=0}^{u} \left(1 - \frac{i}{m}\right), & n \leq m \\ 1 & n > m \end{cases} \tag{8}$$

assuming that all previous users have selected a unique pilot which provides a worst-case analysis. This is converted to pilot collision probability p_{pc}

$$p_{pc}(n,m) = \sum_{u=0}^{n} p_c(u,m) \cdot \frac{u}{n}. \tag{9}$$

Each user transmits a 20 Byte packet and we assume a NOMA rate of 1 Byte per symbol, that results in 20 symbols per user data $\alpha_d = 20$. Non-collided pilots result in successful use of NOMA resource blocks thus we can write $G_{ns} = \alpha_d \cdot n \cdot (1 - p_{pc}(n,m))$. Each NOMA resource block is formed of 4 Physical Resource Blocks (PRBs) each consisting of 12 sub-carriers[1] and 7 symbols. Two of these symbols are used for pilots and 5 of them are used for data transmission. 24 pilots are formed out of these symbols. We set a co-efficient $\alpha_{pi} = \frac{12 \cdot 2 \cdot 4}{24} = 4$ to denote symbols per pilot ratio. This can be translated in to estimation resource in terms of symbols as $G_{ce} = \alpha_{pi} \cdot m$ with m pilots. The NOMA resource allocation is done in constant fashion with 5 symbols in 5 subcarriers and 4 PRBs such that we get $G_{ns} = 12 \cdot 5 \cdot 4$. Through this we write the efficiency as,

$$\rho(n,m) = \frac{\alpha_d \cdot n \cdot (1 - p_{pc}(n,m))}{\alpha_{pi} \cdot m + 240}. \tag{10}$$

Through the settings provided in the paper it boils down to

$$\rho_{\text{PoC}}(n) = \frac{20 \cdot n \cdot (1 - p_{pc}(n,24))}{96 + 240}. \tag{11}$$

2) NORA: We use the same evaluation for the NORA protocol. In LTE a random access block uses 6 PRBs so in total 504 symbols are used for 54 contention based preambles that gives us a $\alpha_p = \frac{504}{54} = 9.3$. This results in $G_{ce} = \alpha_p \cdot m$. We keep the data co-efficient $\alpha_d = 20$ the same to have a comparable approach. There is a protocol difference and the users are allocated resources only after contention is a success. Another improvement is that up to two users accessing the same preamble can be separated thanks to an introduced interference cancellation capability. We will denote this probability as $p_d = 0.6$. With this the success probability $p_s(n)$ for NORA can be re-quoted from [7] as

$$p_s(n,m) = \left(1 + p_d \frac{n-1}{2\,(m-1)}\right) \cdot \left(1 - \frac{1}{m}\right)^{n-1}. \tag{12}$$

Through this equation we have $G_n = 20 \cdot n \cdot p_s(n,54)$. As the NOMA resources are allocated to those users that are successful in the contention phase we have $G_{ns} = G_n$. This can be plugged in Eq. (7) to obtain the efficiency for NORA,

$$\rho_{\text{NORA}}(n) = \frac{20 \cdot n \cdot p_s(n,54)}{504 + 20 \cdot n \cdot p_s(n,54)}. \tag{13}$$

3) MERAP: We can write the G_n in terms of channel estimation pilots m as $G_n = \alpha_d \left(1 - e^{-\lambda_{ce}}\right) m$. The successful NOMa resources is then calculated considering the multiplicity estimation reliability and the channel estimation reliability as in $G_{ns} = \alpha_d \cdot n \cdot p_{ce} \cdot r_{me}$. The m can be calculated given the l using Eq. (6). Using the same co-efficient as PoC we get $G_{ce} = \alpha_{pi} \cdot m$. Last, we use number of maximum users to be supported n_{max} with the constraint Δ_0 to set l. For multiplicity estimation resources l single symbol is required

[1] Each subcarrier is 15 kHz.

Parameter	Explanation	Value
n	Number of users	–
PoC [15]		
G_{me}	Multiplicity estimation resource	0
α_d	Symbols per user data	20
α_{pi}	Symbols per pilot	4
m	Number of pilots	24
NORA [7]		
G_{me}	Multiplicity estimation resource	0
α_d	Symbols per user data	20
α_p	Symbols per preamble	9.3
m	Number of preambles	54
p_d	Preamble interf. canceling probability	0.6
MERAP Proposal		
G_{me}	Multiplicity estimation resource	l
α_d	Symbols per user data	20
α_{pi}	Symbols per pilot	4
m	Number of pilots	vary
r_{me}	Multiplicity estimation ave. reliability	0.9
p_{ce}	Channel estimation reliability	0.9

TABLE II: Summary of parameters used for comparison grouped for different protocols.

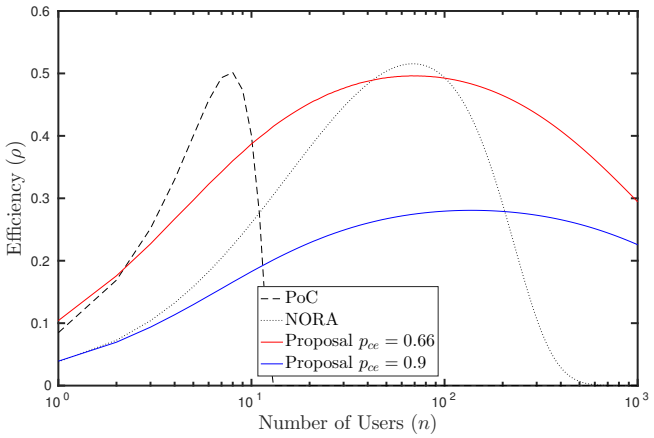

Fig. 7: Efficiency of the proposal is plotted against the state of the art. The Δ_0 is fixed to 10 and $r_{me} = 0.9$ is used.

so the co-efficient is omitted and we get $G_{me} = l$. Then we can plug these results in Eq. (7),

$$\rho_{MERAP}(\Delta_0, n) = \frac{\alpha_d \cdot n \cdot p_{ce} \cdot r_{me}}{l + \alpha_{pi} \cdot m + \alpha_d \left(1 - e^{-\frac{n}{m}}\right) m}. \quad (14)$$

The pilot and NOMA co-efficient is kept as same as the PoC for comparison purposes.

In Fig. 6 we have evaluated how ρ_{mu} behaves with varying Δ_0 for different n values. The behavior is observed to be concave with respect to varying Δ_0 and an optimal dimensioning of the random access based NOMA can be done for a specific arrival rate. However, we want to analyze the behavior of the algorithm for any number of active users.

We have plotted the efficiency versus different number of active users for the proposal on top of the state of the art in Fig. 7. It is seen in the figure that the maximum efficiency is attained when PoC operates around 8 active users per subframe while NORA is designed for operating around 80 active users per subframe. The state of the art solutions focus on certain load i.e., mean arrival and neglect the possible variations in the traffic. Resource efficiency reach a peak of 0.5 and they degrade to 0 out of their operation range. Clearly, against dynamic arrival profiles efficiency will not stay at its peak and system resources will be wasted.

The variance of multiplicity detection is set as $\Delta_0 = 10$ with reliability of $r_{me} = 0.9$ set by using two sigmas. The range of supported number of users with an efficiency of at least 0.3 is increased as expected. However, with high reliability the peak of efficiency decreases to 0.27 compared to the state of the art and also in all regions even though the operation range is extended. Comparable efficiency is reached with lower reliability of $p_{ce} = 0.66$.

As the multiplicity estimator benefits from Poissonization effect, higher efficiency is only enabled with increasing number of users. Other multiplicity estimation techniques such as maximum likelihood can be used to increase the efficiency of lower region. Secondly, the efficiency also decreases with really high number of users since the multiplicity estimator is a single shot estimator. An adaptive multi-round estimator can be used to increase efficiency for high number of users.

VI. Conclusion

In this work we have formulated the resource efficiency problem of the contention based NOMA and evaluated the state of the art solutions. We have seen that both of the solutions are focusing on a certain type of arrival and lack dynamicity against different arrival types. To overcome this we presented a random access protocol for NOMA under MTC paradigm. We believe such a protocol is necessary for an adaptive performance of NOMA against unexpected number of users. The proposed protocol needs a user Multiplicity Estimation step that adjusts the pilot resource allocation. Combination of these two techniques provide exact user multiplicity and non-contaminated channel estimation for NOMA resources. We have also shown a natural trade-off between the multiplicity estimation and channel estimation resources that enables the optimal resource use in the system.

Even though the solution provided has limitations it sheds light to the problems of the contention based use of NOMA resources. We believe that such a view is important for dimensioning of scarce system resources. Future work can improve the estimator to match the efficiency problem of the proposed schemes.

Further work, can include investigating imperfections of the estimation algorithms for user multiplicity detection and the channel estimation in a combined way. Another direction can improve multiplicity estimation with multiple rounds and lower number of symbols.

References

[1] K. N. Pappi, P. D. Diamantoulakis, and G. K. Karagiannidis, "Distributed uplink-NOMA for cloud radio access networks," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2274–2277, Oct. 2017.

[2] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: potential, challenges, and solutions," *IEEE Communications Magazine*, vol. 50, no. 3, 2012.

[3] F. Lazaro and Č. Stefanović, "Finite-length analysis of frameless ALOHA with multi-user detection," *IEEE Communications Letters*, vol. 21, no. 4, pp. 769–772, 2017.

[4] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular iot: Potentials and limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

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

[6] R. E. Learned, A. S. Willsky, and D. M. Boroson, "Low complexity optimal joint detection for oversaturated multiple access communications," *IEEE Transactions on Signal Processing*, vol. 45, no. 1, pp. 113–123, 1997.

[7] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access (NORA) for 5G networks," *IEEE Transactions on Wireless Communications*, 2017.

[8] Y. Du, B. Dong, Z. Chen, X. Wang, Z. Liu, P. Gao, and S. Li, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2812–2828, 2017.

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

[10] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 629–633, May 2016.

[11] E. Björnson, E. De Carvalho, E. G. Larsson, and P. Popovski, "Random access protocol for massive MIMO: Strongest-user collision resolution (SUCR)," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[12] E. De Carvalho, E. Bjornson, J. H. Sorensen, P. Popovski, and E. G. Larsson, "Random access protocols for massive MIMO," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 216–222, 2017.

[13] 3rd Generation Partnership Project (3GPP), "RAN WG2 #71 R2-104663: MTC LTE simulations," Madrid, Aug. 2010, Tech. Rep., 2000. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R2-71--28035.htm

[14] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," *arXiv preprint arXiv:1709.02179*, 2017.

[15] J. Zhang, L. Lu, Y. Sun, Y. Chen, J. Liang, J. Liu, H. Yang, S. Xing, Y. Wu, J. Ma, I. B. F. Murias, and F. J. L. Hernando, "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, June 2017.

[16] M. Kodialam and T. Nandagopal, "Fast and reliable estimation schemes in RFID systems," in *Proceedings of the 12th annual international conference on Mobile computing and networking*. ACM, 2006, pp. 322–333.

[17] L. G. Roberts, "ALOHA packet system with and without slots and capture," *ACM SIGCOMM Computer Communication Review*, vol. 5, no. 2, pp. 28–42, 1975.