A Promising Non-Orthogonal Multiple Access Based Networking Architecture: Motivation, Conception, and Evolution

Dehuan Wan[#], Miaowen Wen[*], *Senior Member, IEEE*,
Xiang Cheng[†], *Senior Member, IEEE*, Shahid Mumtaz[‡], *Senior Member, IEEE*,
and Mohsen Guizani[§], *Fellow, IEEE*

[#]Center for Data Science and Artificial Intelligence
Guangdong University of Finance, Guangzhou 510521, China

[*]School of Electronic and Information Engineering
South China University of Technology, Guangzhou 510641, China

[†]The State Key Laboratory of Advanced Optical Communication Systems and Networks
Department of Electronics, School of Electronics Engineering and Computing Sciences
Peking University, Beijing 100871, China

[‡]Instituto de Telecomunicações
Campus Universitário de Santiago, Aveiro 3810-193, Portugal

[§]Computer Science and Engineering Department
Qatar University, Doha 2713, Qatar

Email: 47-050@gduf.edu.cn, eemwwen@scut.edu.cn, xiangcheng@pku.edu.cn,
smumtaz@av.it.pt, mguizani@ieee.org

Abstract

Since it can offer higher spectral efficiency by granting the served data to share the same spectrum resource synchronously, non-orthogonal multiple access (NOMA) is considered as a bright multiple access technique for future wireless networks. Accommodating a large number of users with NOMA mode may easily cause inappropriate power assignment problem and result in performance degradation. As a result, existing NOMA operation unit mainly focuses on the two-user scenario, where a centre user is only paired with an edge user. However, such a user pairing strategy will also introduce an unbalance to their data rate fairness. Hence, it is of great importance to develop an effective structure that facilitates low decoding complexity, high system performance, and reasonable user fairness with NOMA. To this end, this article proposes a novel NOMA operation unit, in which successive interference cancellation can be performed to accommodate an arbitrary number of users by integrating NOMA technique with the orthogonal resource allocation strategy. Besides, the proposed NOMA operation unit can further improve edge user's achievable rate by incorporating the index modulation technique according to the data service demanded by centre users.

I. Introduction

With the rapid development of wireless transmission and mobile networking technologies, there is an explosive growth in both the quantity and diversity of wireless services and smart devices. To support the huge demand on data traffic, non-orthogonal multiple access (NOMA) is considered as a bright multiple access technique for future mobile communications systems [1], [2]. It can be primarily realized in code and power domains [1], and combined with either single-carrier or multi-carrier transmission [2]. Although multi-carrier NOMA can promise attractive advantages, there exist some challenges in such as sophisticated codebook design and trading off the system performance and complexity [3]. On the other hand, thanks to its implementation flexibility, power domain NOMA has been applied in various scenarios.

NOMA breaks the limit of the existing orthogonal multiple access (OMA) techniques, and exhibits huge inherent potential in the following aspects:

- Higher capacity: The reason lies in that NOMA can share the entire time and frequency resources for all served users by superimposing their signals with different powers, whereas the OMA users are usually limited by the degree of freedom in time or frequency.
- Denser connection: Unlike OMA, in which the principle of orthogonality has put strict limitation on the number of connections, NOMA can in theory accommodate an arbitrary number of users by superimposing their signals regardless of the resource orthogonality.
- Lower latency: By contrast with 10 ms required by the 3rd Generation Partnership Project (3GPP) Long-Term Evolution (LTE), NOMA is expected to offer lower latency due to its flexible user scheduling.

Besides, NOMA shows good compatibility with other existing techniques, such as massive multiple-input multiple-output (MIMO), millimeter wave (mmWave) communications [4], and wavelet fourier transform [5].

For NOMA, a massive access of the served users synchronously may easily result in compromised system performance with inappropriate power allocation. As a result, the current NOMA operation unit mainly focuses on the two-user scenario, where a centre user is paired with an edge user [6]–[8]. Such a two-user pairing strategy can reduce the decoding complexity of the centre user who has limited processing capability [9]. However, it may also result in poor performance/fairness for the edge user with respect to the centre one. It has been pointed out

that by allocating more power to the edge user via, e.g., optimizing the minimum rate of the paired users [10] or ensuring the rates of both users to be the same as that provided by OMA [11], their performance/fairness can be improved significantly. Unfortunately, by doing so, the system's sum rate will decrease since the power assigned to the centre user is less. On the other hand, the superiority of NOMA over OMA is highly dependent on the near-far effect constructed by the involved users [9]. The existing pairing strategy that considers only two users cannot necessarily exploit the advantages brought by NOMA. Configuring or selecting the involved users, not necessarily two, to further exploit the near-far effect of the considered network is a goal of this article.

The rest of this article is arranged as follows. We first give an overview of the literature on the existing NOMA operation units, and then describe the proposed NOMA operation unit and its generalized form, where the three scheduled users case is presented to illustrate the advantages in Section II. The proposed operation unit with index modulation (IM), which is developed to further improve the edge user's performance, is illustrated in Section III. In Section IV, the main challenges on network implementation are discussed. Finally, Section V concludes this article.

II. Structures of NOMA Operation Cell

A. The Existing NOMA Scheme

Assume that the number of users is L, and channel coefficients between the base station and L users $\{U_1, \ldots, U_L\}$ are independent of each other with average channel gains $\{\gamma_{U_1}, \ldots, \gamma_{U_L}\}$, satisfying $\gamma_{U_1} \leq \cdots \leq \gamma_{U_L}$. An effective power allocation scheme can be employed, where the power allocation coefficient associated with U_l is $p_l = 2\frac{L^2+2L+2-(L+2)l}{L(L^2+L+2)}$, which is different from the one proposed in [6], where $p_1 = 4/5$ and $p_2 = 1/5$ for $L = 2$ and $p_l = 2\frac{L-l+1}{L(L+1)}$ with $\sum_{l=1}^{L} p_l = 1$ for $L > 2$.[1] However, it is inadvisable to serve all users in a resource unit to perform NOMA. The reasons are twofold. Firstly, such a network is an interference-limited system, and accommodating a large number of users may easily cause inappropriate power assignment problem and result in performance degradation. A $L = 3$ case is shown in Fig. 1,

[1]The dynamic power allocation coefficients according to instantaneous channel state information (CSI) can also be applied to the considered network. For the sake of presentation, we assume the statistical S-CSI is known to the considered network, and the fixed power allocation scheme is naturally designed for the proposed system.

from which we can see the outage probability with the proposed power allocation factors is better than that with the power allocation factors proposed in [6] and that with random power allocation. Secondly, channel gain differences cannot be guaranteed to be distinct enough. To solve these problems, users are separated into different groups and radio resources are shared non-orthogonally within each group and orthogonally among different groups.[2] A typical two-user pairing NOMA operation unit is investigated mathematically in [6], where the power allocation factors for cell-edge user (U_E) and cell-centre user (U_C) are fixed as $p_1 = 4/5$ and $p_2 = 1/5$, respectively, and they both share the entire bandwidth B. In fact, the superiority of NOMA over OMA can be further improved by scheduling the users who show distinct channel conditions appropriately. This can be realized by using the distributed matching algorithm [7], which is shown to improve system performance significantly. However, on the one hand, the scheduling algorithm is complex, especially when users' number is large, e.g., in a macro-cell scenario; on the other hand, even using this algorithm, users may not be optimally paired if the search is limited to a single unit or the OMA operation among groups is based on time and/or frequency. Moreover, as shown in Fig. 2(a), the fairness of such a two-user NOMA scheme in terms of user's achievable rate is lower than that of OMA at high signal-to-noise ratio (SNR), in which the resources allocated to both users are non-overlapped. This can be explained by the fact that the edge user takes the centre user's signal as interference signal to decode its symbol resulting in a fixed achievable rate $\log_2(1 + \frac{p_1}{p_2})$ with $\rho \to \infty$, while the achievable rate of the centre user can be approximated as $\log_2(\rho)$, where ρ denotes the SNR. The fairness can be improved if power coefficients are allocated appropriately to satisfy $p_1 \gg p_2$, but such power allocation strategy results in system sum rate degradation since the power assigned to the centre user who has better channel gain is very small, as shown in Fig. 2(b).

B. Proposed Structure

1) Concept and Design: For the conventional NOMA scheme, channel gains of the grouped users in the same operation unit are necessarily distinct, and can be sorted in an ascending or a descending order. Besides, only two scheduled users are multiplexed to reduce the complexity

[2]For multicarrier NOMA, such as sparse code multiple access (SCMA), no scheduled user can share all subcarriers. However, in our proposed scheme, paired with each cell-centre user to perform NOMA the cell-edge user shares the whole bandwidth.

caused by the application of successive interference cancellation (SIC). However, as mentioned earlier, the fairness of the conventional scheme is poor at high SNR [6]. Although some approaches have been suggested to guarantee the quality of service (QoS) of the edge user, such as the achievable rate or outage probability [10], [11], the problem of user fairness has not been satisfactorily solved. In the sequel, we will answer the following questions:

- How to perform NOMA effectively when there exist many centre users whose channel conditions are not distinct enough?
- Can we enhance user fairness?

Consider a cell network, which is comprised of one base station (BS), M cell-edge users and K cell-centre users with $M < K$. We assume that the K cell-centre users can be divided into M groups to form a multiple-unit cell, of which the m-th operation unit has a cell-edge user U_{mE} and N^m cell-centre users $\{U_{m1}, \ldots, U_{mN^m}\}$ with $\sum_{m=1}^{M} N^m = K$, and the N^m cell-centre users have close average channel gains, as shown in Fig. 3(a). The orthogonal frequency division strategy is applied to the N^m cell-centre users of the m-th operation unit, namely there are N^m orthogonal subcarriers serving the $N^m + 1$ users, and thus the cell-edge user U_{mE} is paired with each cell-centre user to form a multiple-pairing unit to perform NOMA, as shown in Fig. 3(b). More specifically, assuming that the total bandwidth for the m-th group is B, the frequency band assigned to U_{mi} (which denotes the i-th cell-centre user in the m-th group with $i \in \{1, \ldots, N^m\}$) is $B_i = B/N^m$, and U_{mE} is paired with each U_{mi} to perform conventional NOMA. This means that U_{mE} will be paired N^m times. Let U_{mEi} denote the image of U_{mE} that is paired with U_{mi} in the i-th NOMA operation pair. Therefore, for the i-th pair, the power allocation coefficients for U_{mi} and U_{mEi} are $\alpha_i = p_2/N^m$ and $\beta_i = p_1/N^m$ according to the conventional two users structure, where $p_1 = 3/4$ and $p_2 = 1/4$ from our proposed power allocation scheme, and the total power and bandwidth allocated to U_{mE} amount to p_2 and B, respectively. Clearly, such a grouping scheme will not bring heavy computational burden on U_{mi} to perform SIC as only two users are scheduled to do NOMA. In addition, it is costless for the cell-edge user to obtain its required signals as the decoding in different frequency bands can be performed in parallel. The frequency bands can be flexibly configured to either boost the data rate given an outage probability to harvest the multiplexing gain or improve the outage performance given a data rate

to harvest the diversity gain.[3] This property is shown in Fig. 4.

2) Performance demonstration: To better illustrate the advantages of our proposed NOMA operation unit, we conduct numerical analysis with $N^m = 2$ for the m-th group, where there exist two U_{mi}, $i = 1, 2$, and U_{mE}, and $\alpha_i = 1/8$ and $\beta_i = 3/8$ for the i-th pair. To begin with, let us take a brief overview of the conventional NOMA and OMA for a three-user case. Specifically, for conventional NOMA, more power is allocated to the user requiring higher QoS as the channel qualities for the two cell-centre users are almost identical. Assume that U_{m2} has a higher QoS than U_{m1}, and the power allocation coefficients for U_{mE}, U_{m2}, and U_{m1} are $p_3 = 3/6$, $p_2 = 2/6$, and $p_1 = 1/6$, as suggested in [6]. For OMA, we take frequency division multiple access (FDMA) and time division multiple access (TDMA) as examples, where the radio resources are averagely assigned to all involved users. As performance metrics, we take the sum rate, outage behavior, and fairness. Detailed simulation results are provided in Fig. 4, where one can see that TDMA shows worse performance than FDMA due to the fact that TDMA receives background noise across the entire bandwidth, and the proposed NOMA scheme also significantly outperforms TDMA. However, compared with FDMA, the proposed NOMA exhibits superiority on achievable sum rate but inferiority on outage probability if the cell-edge user wants to acquire full multiplexing, as shown in Fig. 4(a), or better performance on both achievable sum rate and outage probability if the cell-edge user wants to acquire full diversity, as shown in Fig. 4(b). In fact, the outage performance of the cell-edge user with full multiplexing can be further improved and even better than its counterparts in FDMA if more power is assigned to it. It can be understood as the multi-user interference suffered by the cell-edge user from the cell-centre user will be reduced. Moreover, the proposed NOMA overwhelms the conventional NOMA in terms of outage performance and fairness as the former can suppress the multi-user interference effectively, and they can be comparable in terms of achievable sum rate when the cell-edge user acquires full multiplexing.

C. Generalized Proposed NOMA Operation Unit

In the m-th NOMA operation unit, as shown in Fig. 3(b), U_{mE} is paired with N^m centre users to implement NOMA. Therefore, there exist $C(N^m, M)$ combinations for U_{mE} to select N^m

[3]The full multiplexing here denotes that the cell-edge user's required signals carried by each frequency band are different; while the full diversity denotes that the cell-edge user's required data transmitted by different frequency band are the same.

cell-centre users to perform NOMA, which is different from conventional NOMA whose user scheduling is only limited to a local cluster determined by coordinated beamforming [12]. We can find the global optimal user scheduling scheme for the cell-edge user to further improve its performance by selecting the cell-centre users according to the given objective function, such as sum rate, outage probability, or user fairness, with the existing practical algorithms, e.g., monotonic optimization, successive convex approximation, and game theory [13].

In fact, it is not necessary to guarantee that all the N^m cell-centre users who perform NOMA with U_{mE} in the m-th group have close average channel gains. Consider the case that the cell-center users in the m-th group have different average channel gains, where the fluctuation of the average channel power among them is not big, and $\gamma_{U_{m1}} \leq \cdots \leq \gamma_{U_{mN^m}}$. For the generalized architecture of the proposed NOMA operation unit, a simple and effective power allocation for the i-th NOMA operation pair in the m-th group can be obtained as $\alpha_i = w_1 \frac{\gamma_{U_{mi}}}{\sum_{i=1}^{N^m} \gamma_{U_{mi}}}$ and $\beta_i = w_2 \frac{\gamma_{U_{mi}}}{\sum_{i=1}^{N^m} \gamma_{U_{mi}}}$, respectively, where w_1 and w_2 are the weights for the center and edge users, respectively, with $w_1 + w_2 = 1$. A special case is that the weights can be set as $w_1 = 1/4$ and $w_2 = 3/4$ according to the power allocation coefficient proposed in Section II.A for $N = 2$. Finally, the scheduled users in the proposed generalized model can be allocated with different powers, depending on their channel conditions.

III. The proposed NOMA Operation Unit with IM technique

IM technique prevails for its capability of striking interesting tradeoff between the spectral efficiency and energy efficiency by delivering the activation states of some radio resources of the communication systems to realize information embedding [14]. By integrating NOMA with IM techniques via the data service demanded of the centre users, the proposed NOMA operation unit can further improve the edge user's performance.

Generally, a data service demanded by the user can be usually classified as either real-time (RT) or non-real-time (NRT). A RT service requires guaranteed latency while the NRT service shares the total system capacity. Therefore, RT service is given a strictly higher priority than NRT, which is consistent with the existing studies that a cell-centre user is usually viewed as a secondary user (SU) while the cell-edge user is regarded as the primary user (PU) [15]. Inspired by the IM [14] concept and with the demand of RT and NRT services of the cell-centre users, the proposed NOMA operation unit can be developed to an IM mode (IM-NOMA).

Without loss of generality, assume that $N^m = 4$, where U_{m1} and U_{m2} require RT services while U_{m3} and U_{m4} demand NRT services, as shown in Fig. 5(a). According to the analysis in Section II, the bandwidths for the pairs of U_{m1} and U_{mE}, U_{m2} and U_{mE}, U_{m3} and U_{mE}, and U_{m4} and U_{mE} are B_1, B_2, B_3, and B_4, respectively. The two RT service users are activated to pair with U_{mE} during each transmission, whereas only one of the two NRT service users, either U_{m3} or U_{m4}, is activated to pair with U_{mE}. The active NRT service user is selected by the input bits of U_{mE}. In this regard, U_{mE}'s input bits can be divided into two parts: the transmit bits d_{mEi} at each NOMA operation pair and the index bits d^I_{mE} for activating the NRT service user. Specifically, we use $d^I_{mE} = 0$ to represent the case that U_{m3} is selected as the activated user while $d^I_{mE} = 1$ for U_{m4} and vice versa, as shown in Fig. 5(b). This rule is pre-known by U_{mE}, and the bit message for selecting the active NRT service users can be decoded by U_{mE} via active carrier frequency sensing. A description about the above processing is also presented in Fig. 6(a), where the cell-edge user employs binary phase-shift keying (BPSK) with input data $d_{mE1} = 1$, $d_{mE2} = 1$, $d_{mE3} = 0$, and $d^I_{mE} = 1$ while the cell-centre users use four-quadrature amplitude modulation (4-QAM) with input data $d_{m1} = 01$, $d_{m2} = 00$, $d_{m3} = 11$, and $d_{m4} = 10$. As shown in Fig. 6(b), the edge user's achievable rate with the IM technique can be further improved.

IV. Challenges on Network Implementation

As a new network architecture, although it can enhance the spectral efficiency and cell-edge user's performance significantly, there exist some open issues and challenges, especially on users' scheduling and their resource management.

A. Feedback Overhead and Users' Scheduling Complexity

It has been pointed out that we can find the global optimal user scheduling scheme for edge user to further improve its performance via feedback signaling. Although the base station has a centralized processing unit and shows excellent processing capacity, a non-negligible overhead may be caused unavoidably, especially for the scenario with with huge number of high mobility users. Fortunately, the existing studies support that NOMA is able to outperform OMA with statistical channel state information (CSI). Therefore, for the low mobility cases, channel feedback will be relaxed due to the fact that the average channel gains can be viewed as

an effective parameter for user scheduling. In this regard, the requirement of precise CSI at the base station for multiplexing the users via power allocation will be relaxed, which will reduce feedback overhead dramatically.

B. Resource Management Based on CSI

NOMA allows the multiplexed users to enjoy the radio resources synchronously but different power factors are assigned to them based on their channel gains. Thus, for conventional NOMA operation unit, first, the channel differences among the users must be distinct, and then an effective power allocation scheme must be designed for the involved users with their available CSI knowledges; otherwise the superiority of NOMA over OMA might diminish. However, for the proposed NOMA operation unit, paired with each cell-center user whose frequency band is allocated orthogonally, the cell-edge user performs NOMA and shares the total bandwidth. As such, user fairness is improved significantly compared to the conventional NOMA operation unit. Moreover, if a flexible power allocation scheme, e.g., the adaptive power allocation strategy, can be also designed for the proposed network, system performance will be further improved, especially when such scheme is determined by the precise CSI. However, when network impairment such as channel aging, feedback delay, or some abrupt inter-cell interference happens, which is a common case in practical networks, precise CSI will not be available and thus system's performance will be degraded accordingly. From the aforementioned discussion, for low mobility cases, a practical power allocation scheme should be designed using the statistical CSI, namely the average channel gains.

V. Conclusions

In this article, the weakness of the existing single-cell NOMA operation unit has been discussed from the perspective of user fairness, and a promising alternative has been proposed. A generalized single-group network architecture constructed by the proposed NOMA operation unit has been designed, where the globally optimum user scheduling scheme can be found to further improve system's performance. The generalized network architecture has been further developed by incorporating the IM technique according to the data service demanded by cell-centre users. It has been pointed out that such an IM-NOMA network can further improve cell-edge user's

achievable rate. Finally, the key challenges for the design of network architecture with the proposed NOMA operation unit have also been highlighted.

VI. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 61871190, in part by the Natural Science Foundation of Guangdong Province under Grant 2018B030306005, and in part by the Pearl River Nova Program of Guangzhou under Grant 201806010171. The work has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 815178.

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[2] 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.

[3] M. Moltafet, N. M. Yamchi, M. R. Javan, and P. Azmi, "Comparison study between PD-NOMA and SCMA," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1830–1834, Feb. 2018.

[4] D. Zhang, Z. Zhou, C. Xu, Y. Zhang, J. Rodriguez, and T. Sato, "Capacity analysis of NOMA with mmWave massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1606–1618, July 2017.

[5] S. Baig, U. Ali, H. M. Asif, A. A. Khan and S. Mumtaz, "Closed-form BER expression for fourier and wavelet transform based pulse-shaped data in downlink NOMA," *IEEE Commun. Lett.*, doi: 10.1109/LCOMM.2019.2903083, 2019.

[6] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 1462–1465, Aug. 2016.

[7] 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.

[8] A. Zafar, M. Shaqfeh, M. S. Alouini, and H. Alnuweiri, "On multiple users scheduling using superposition coding over Rayleigh fading channels," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 733–736, Apr. 2013.

[9] 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 Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, Oct. 2017.

[10] X. Liu and H. Jafarkhani, "Downlink non-orthogonal multiple access with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6151–6164, Sept. 2017.

[11] J. A. Oviedo and H. R. Sadjadpour, "A fair power allocation approach to NOMA in multi-user SISO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7974–7985, Sept. 2017.

[12] Z. Chen, Z. Ding and X. Dai, "Beamforming for combating inter-cluster and intra-cluster interference in hybrid NOMA systems," *IEEE Access*, vol. 4, pp. 4452–4463, 2016.

[13] 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.

[14] M. Wen, X. Cheng, L. Yang, *Index Modulation for 5G Wireless Communications*. Berlin, Germany: Springer, 2017.

[15] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10152–10157, Dec. 2016.

BIOGRAPHIES

Dehuan Wan (47-050@gduf.edu.cn) received the Ph.D. degree from South China University of Technology, Guangzhou, China, in 2018. He is currently a lecturer with the Guangdong University of Finance, Guangzhou, China. His research interests include non-orthogonal multiple access, index modulation, and data science and artificial intelligence.

Miaowen Wen [SM'18] (eemwwen@scut.edu.cn) received his Ph.D. degree in Signal and Information Processing from Peking University, China, in 2014. From 2012 to 2013, he was a Visiting Student Research Collaborator with Princeton University, USA. He is currently an Associate Professor with South China University of Technology, China, and a Post-doctoral Fellow with The University of Hong Kong, China. He has authored a book and more than 60 IEEE journal papers. He was the recipient of Best Paper Awards at IEEE ITST'12, ITSC'14, and ICNC'16. He served as a Guest Editor for IEEE JSAC, IEEE JSTSP, and IEEE Access. His research interests include index modulation, non-orthogonal multiple access, physical layer security, and molecular communications.

Xiang Cheng [SM'13] (xiangcheng@pku.edu.cn) received his Ph.D. degree from Heriot-Watt University and the University of Edinburgh, United Kingdom, in 2009, where he received the Postgraduate Research Thesis Prize. He is currently a professor at Peking University, Beijing, China. His general research interests are in the areas of channel modeling and communications. He was the recipient of the IEEE Asia Pacific (AP) Outstanding Young Researcher Award in 2015, and Best Paper Awards at IEEE ITST'12, ICCC'13, ITSC'14, ICC'16, and ICNC'17. He has served as Symposium Leading-Chair, Co-Chair, and a member of the Technical Program Committee for several international conferences. He is now an Associate Editor for *IEEE Transactions on Intelligent Transportation Systems*.

Shahid Mumtaz [SM'16] (smumtaz@av.it.pt) is an ACM Distinguished speaker, IEEE Senior member, EiC of IET "journal of Quantum communication" Vice Chair: Europe/Africa Region-

IEEE ComSoc: Green Communications & Computing society and Vice-chair for IEEE standard on P1932.1: Standard for Licensed/Unlicensed Spectrum Interoperability in Wireless Mobile Networks. He has more than 12 years of wireless industry/academic experience. He has received his Master and Ph.D. degrees in Electrical & Electronic Engineering from Blekinge Institute of Technology, Sweden, and University of Aveiro, Portugal in 2006 and 2011, respectively. He has been with Instituto de Telecomunicações since 2011 where he currently holds the position of Auxiliary Researcher and adjunct positions with several universities across the Europe-Asian Region. He is also a visiting researcher at Nokia Bell labs. He is the author of 4 technical books, 12 book chapters, 150+ technical papers in the area of mobile communications.

Mohsen Guizani [S'85, M'89, SM'99, F'09] (mguizani@ieee.org) received the B.S. (with distinction) and M.S. degrees in electrical engineering, and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently a professor in the CSE Department at Qatar University, Qatar. He is an IEEE Fellow and a Senior Member of ACM.

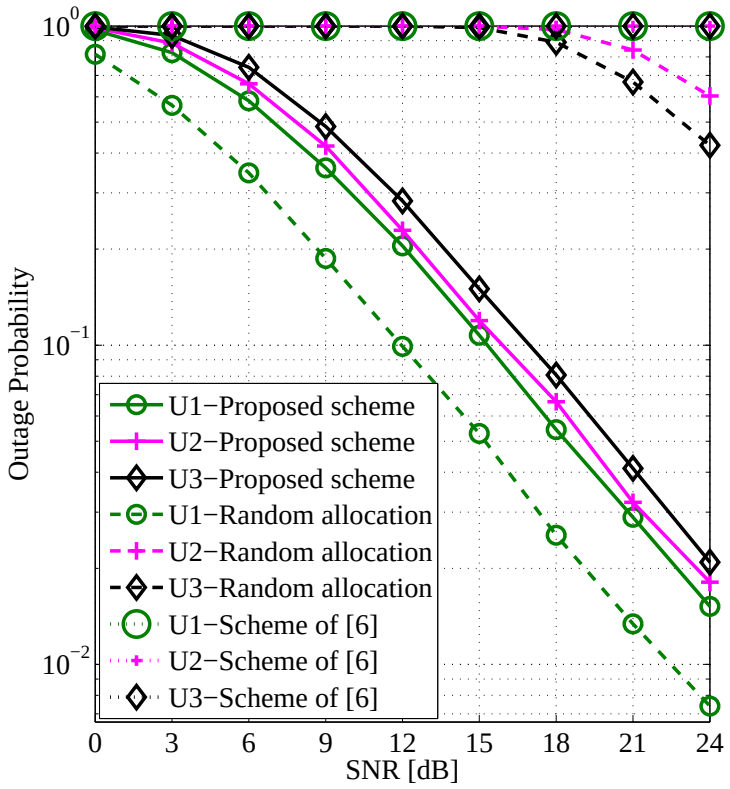


Fig. 1: Outage probability of the single-cell NOMA network with user 1 (U_1), user 2 (U_2), and user 3 (U_3) under different power allocation coefficients, where: (a) $p_1 = 0.65$, $p_2 = 0.25$, and $p_3 = 0.1$ from random fixed power allocation scheme (dashed lines); (b) $p_1 = 12/21$, $p_2 = 7/21$, and $p_3 = 2/21$ from our proposed power allocation scheme (solid lines); (c) $p_1 = 3/6$, $p_2 = 2/6$, and $p_3 = 1/6$ from [6] (dotted lines). The channel settings for U_1, U_2, and U_3 are $\gamma_{U_1} = 2$, $\gamma_{U_2} = 6$, and $\gamma_{U_3} = 10$, respectively, and the rate thresholds for them are $R_{th}^1 = 0.5$ bps/Hz, $R_{th}^2 = 0.9$ bps/Hz, and $R_{th}^3 = 1.3$ bps/Hz, respectively.

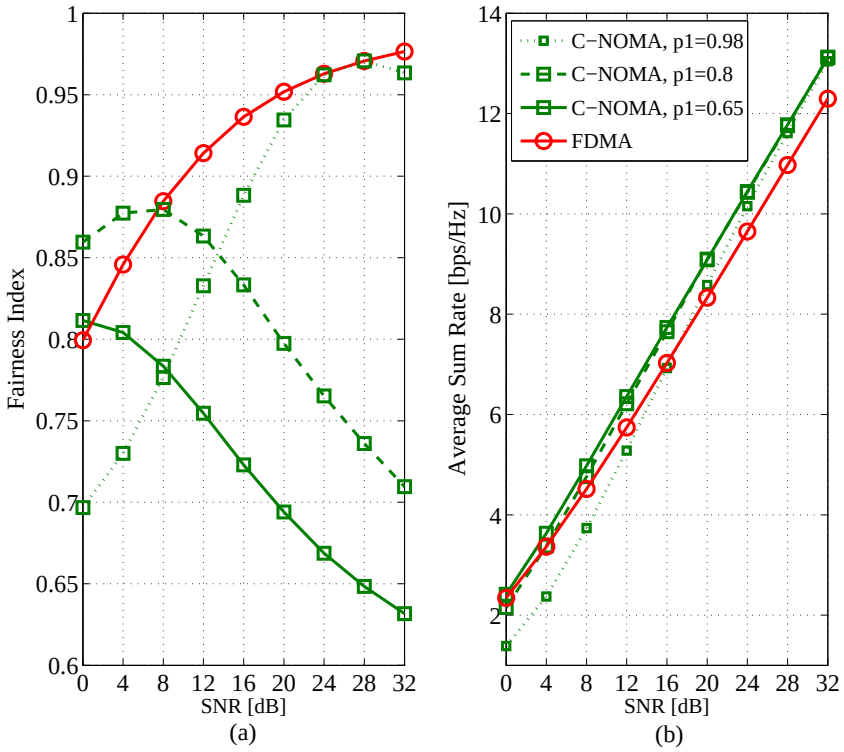


Fig. 2: Comparison of Jains Index (a) and sum rate (b) between NOMA and OMA under different power allocation coefficients, where FDMA is selected as a representative for OMA. The channel settings for U_E and U_C are $\gamma_{U_C} = 10$ and $\gamma_{U_E} = 2$, respectively. The power allocation coefficients for U_E and U_C with NOMA are: (1) $p_1 = 0.65$ and $p_2 = 0.35$ (solid lines); (2) $p_1 = 0.8$ and $p_2 = 0.2$ (dashed lines); (3) $p_1 = 0.98$ and $p_2 = 0.02$ (dotted lines), while the resources allocated to U_E and U_C in FDMA are equal.

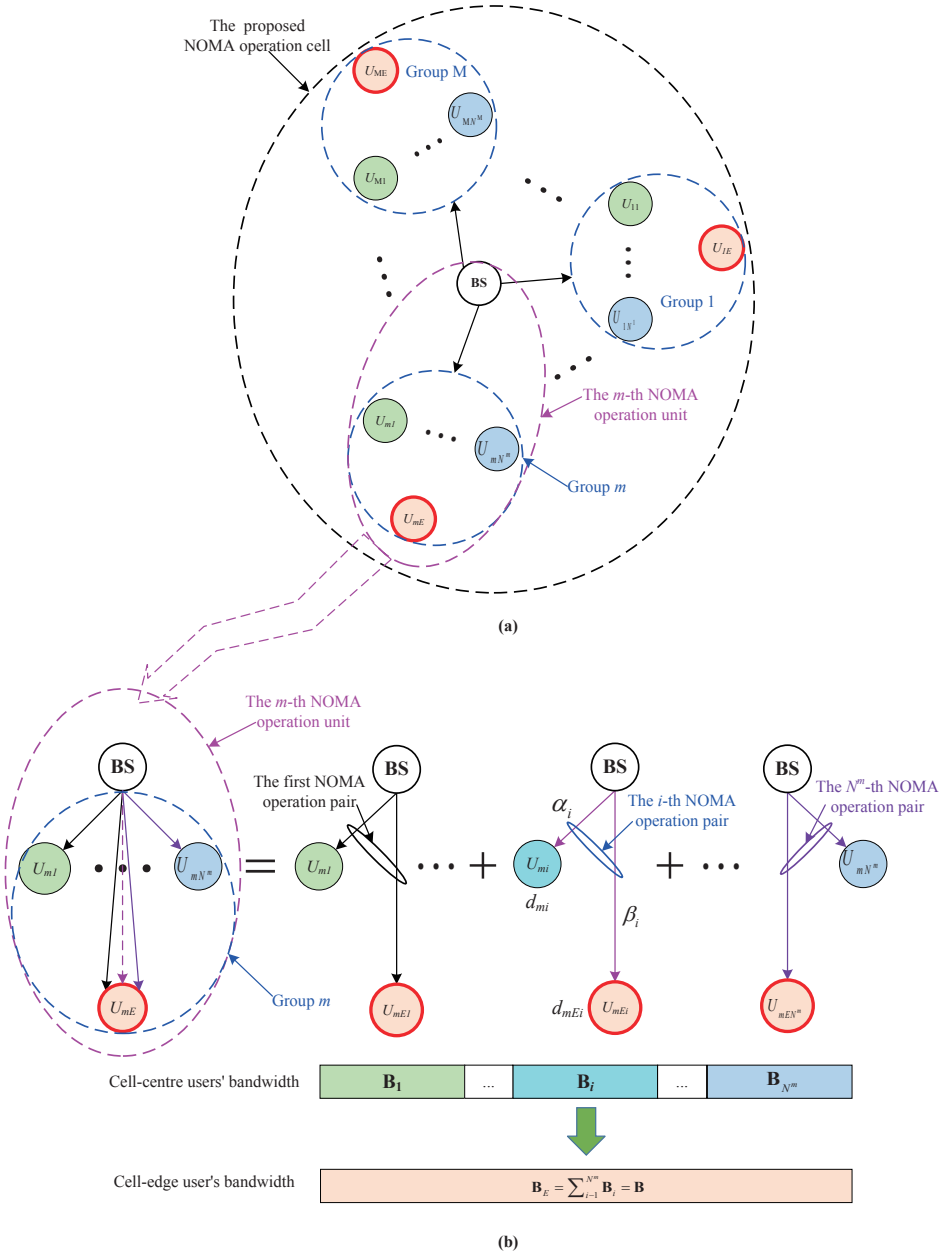


Fig. 3: Proposed network structure: (a) the NOMA operation cell comprised of M operation units; (b) the m-th NOMA operation unit comprised of P^m operation pairs, and the bandwidth orthogonally allocated to each operation pair.

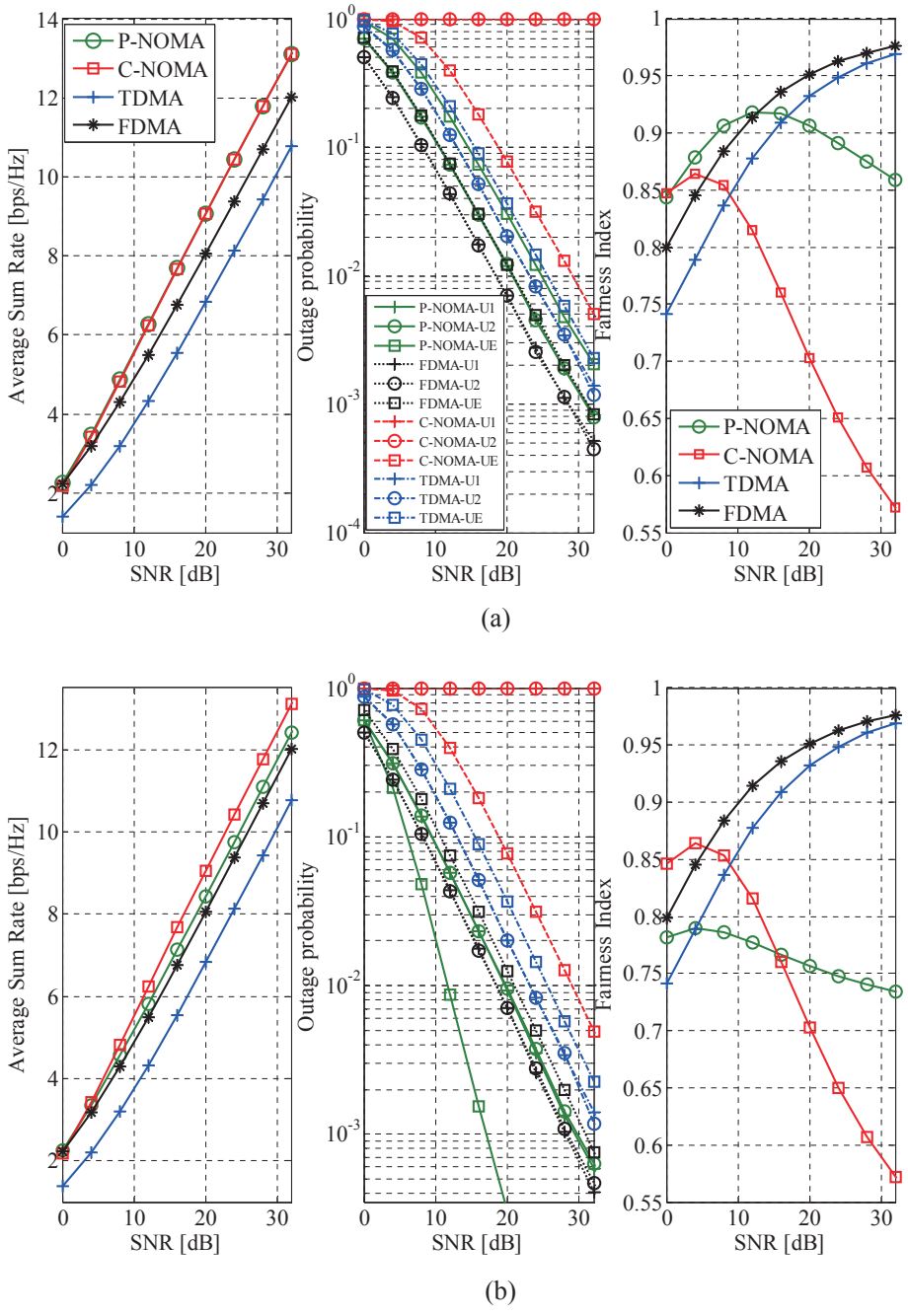
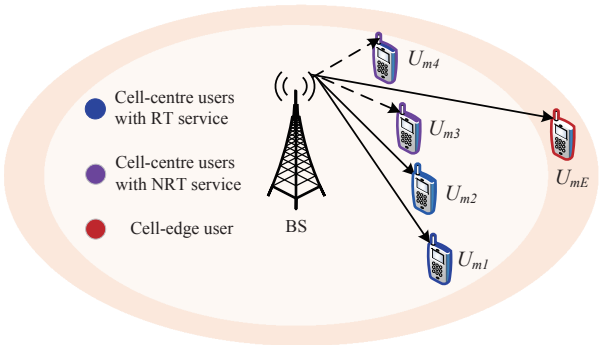


Fig. 4: Comparison of the achievable sum rates, outage probability of individual user, and fairness with (a) full multiplexing gain or (b) full diversity gain for cell-edge user among the proposed NOMA (P-NOMA), conventional NOMA (C-NOMA), and OMA schemes. The channel settings for U_1, U_2, and U_E are $\gamma_{U_1} = 10$, $\gamma_{U_2} = 10$, and $\gamma_{U_E} = 2$, respectively. The rate thresholds for U_1, U_2, and U_E are $R_{th}^1 = 1$ bps/Hz, $R_{th}^2 = 1$ bps/Hz, and $R_{th}^E = 0.6$ bps/Hz, respectively.

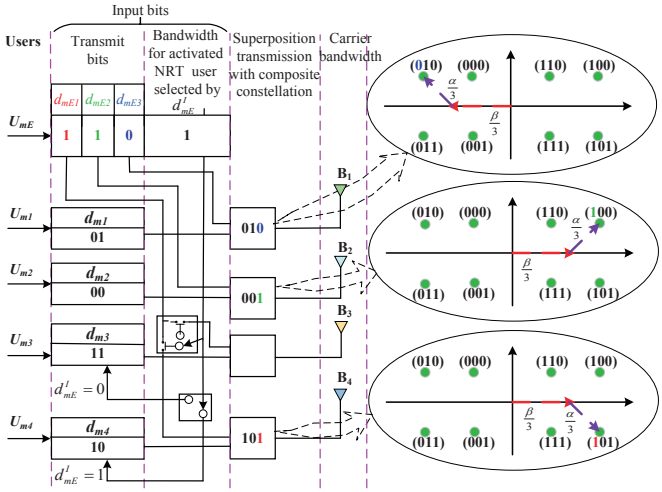


(a)

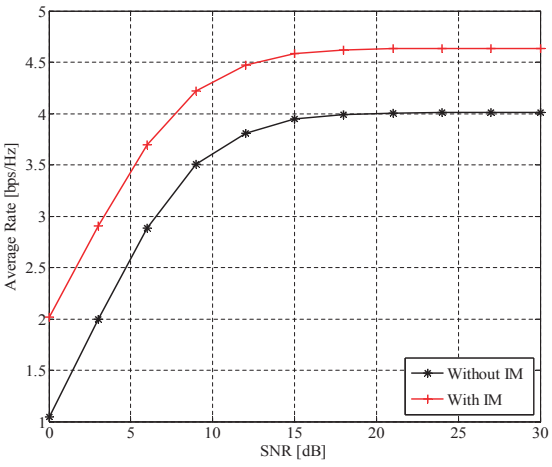
		Bandwidths allocated for the activated RT users		Bandwidth allocated for the activated NRT user (selected by index bit d^I_{mE})	
Total bandwidths for each transmission	$d^I_{mE}=0$	B_1	B_2	B_3	
	$d^I_{mE}=1$	B_1	B_2		B_4
Transmitted bits in the corresponding bandwidths	$d^I_{mE}=0$	d_{m1} and d_{mE1}	d_{m2} and d_{mE2}	d_{m3} and d_{mE3}	
	$d^I_{mE}=1$	d_{m1} and d_{mE1}	d_{m2} and d_{mE2}		d_{m4} and d_{mE3}
Users' input bits for each transmission	$d^I_{mE}=0$	$\underbrace{d_{m1}}_{U_{m1}\text{'s}}+\underbrace{d_{m2}}_{U_{m2}\text{'s}}+\underbrace{d_{m3}}_{U_{m3}\text{'s}}+\underbrace{d_{mE1}+d_{mE2}+d_{mE3}+d_{mEI}}_{U_{mE}\text{'s}}$			
	$d^I_{mE}=1$	$\underbrace{d_{m1}}_{U_{m1}\text{'s}}+\underbrace{d_{m2}}_{U_{m2}\text{'s}}+\underbrace{d_{m4}}_{U_{m4}\text{'s}}+\underbrace{d_{mE1}+d_{mE2}+d_{mE3}+d_{mEI}}_{U_{mE}\text{'s}}$			

(b)

Fig. 5: The proposed NOMA operation unit for the m-th group: (a) a generalized architecture; (b) an IM architecture with $N^m = 4$; (c) the IM principle of (b).



(a)



(b)

Fig. 6: The proposed IM-NOMA operation unit for Fig. 5(b): (a) composite constellation with superposition transmission; (b) achievable rate of the cell-edge user achieved by the proposed scheme with or without IM technique.