Resource Allocation in 5G with NOMA-Based Mixed Numerology Systems

Ru-Jun Wang†, Chih-Hang Wang♯, Guang-Siang Lee♯, De-Nian Yang♯,
Wen-Tsuen Chen†, and Jang-Ping Sheu†
†Dept. of Computer Science, National Tsing Hua University, Taiwan
♯Institute of Information Science, Academia Sinica, Taiwan
E-mail: s107062649@m107.nthu.edu.tw, {superwch7805, gslee9822802, dnyang}@iis.sinica.edu.tw,
and {wtchen, sheujp}@cs.nthu.edu.tw

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322172

***Abstract*—New radio (NR) and non-orthogonal multiple access (NOMA) have emerged for more scalable and efficient resource utilization in 5G. NR implements mixed numerology with a flexible radio frame structure to ensure forward compatibility for future services, whereas NOMA allows multiple users with different channel states to share identical radio resources. However, the resource allocation in the NOMA-based mixed numerology system is challenging due to the naturally different shapes of Physical Resource Block (PRB) for NR and the reused locations of PRBs in a radio frame for NOMA. In this paper, we formulate a new optimization problem Multi-Dimensional Resource Allocation Problem (MDRAP) and prove that MDRAP is NP-hard. To solve the problem, we propose an approximation algorithm to maximize the weighted sum rate under the heterogeneity of users. The algorithm includes *Zone Displacement* to displace the locations of allocated PRBs in different layers of the radio frame, and *Zone Allocation* to change the location of the bounded rectangles (i.e., zones) for the allocation in each layer. We design *Layer Dissimilarity* to examine the location and shape of PRBs for avoiding inter-numerology interference between different layers. Simulation results show that the proposed algorithm outperforms state-of-the-art algorithms regarding throughput and fairness.**

I. Introduction

The fifth-generation (5G) network is projected to support massive wireless connections with diverse traffic demands. To this end, 3GPP recently specified the *New Radio* (NR) access technology to allocate Physical Resource Blocks (PRBs) in a more flexible way. NR implements multi-numerology with a flexible radio frame structure that ensures 5G forward compatibility along with the adjustable subcarrier spacing (SCS) and transmission time interval (TTI) [1]. Different from 4G LTE, there are multiple *shapes*[1] of basic scheduling unit (i.e., PRB) in 5G for diverse user demands [1]. For example, the PRBs with shorter length[2] in the frequency domain have small bandwidth and are appropriate for non-delay-sensitive and low volume data transmission like environmental sensing. In contrast, the PRBs with longer length but shorter width are suitable for services with stringent latency demands like vehicle-to-vehicle communications [2], [3]. Therefore, the PRB allocation becomes more complicated since the available numerologies are distinct for each user, and different shapes of PRBs may not be well aligned in a radio frame and lead to waste PRBs.

Non-orthogonal multiple access (NOMA) is another crucial technology for increasing the spectrum efficiency in 5G networks. Different from the traditional OMA in 4G, where a single PRB can be allocated to only one user [4], NOMA allows multiple users with different channel states to multiplex the identical PRB in the code or power-domain [2], [5]. By applying the superposition coding (SC) and successive interference cancellation (SIC), users with higher channel gains can successfully decode the data and remove the inter-user interference from the users with lower channel gains [5], and the system throughput can be improved by reusing identical spectrum resource [4]. Although NOMA can null inter-user interference, spectrum sharing with different numerologies will suffer from the Inter-Numerology Interference (INI) attributed to misalignment in receiving window length [6]. Compared to the PRB allocation in 4G LTE, it is more challenging in a NOMA-based mixed numerology system due to the diverse shapes of PRBs (i.e., different numerologies result in different PRB shapes) and the effect of interference on PRB reusage.

Previous research on the NOMA-based mixed numerology system mainly focused on spectrum sharing [6]–[8]. Popovski *et al*. [7] designed a communication-theoretic model for spectrum sharing between heterogeneous services. Choi *et al*. [6] derived the INI pattern in the mixed numerology spectrum sharing system and proposed to restrain the growing of interference. McWade *et al*. [8] analyzed the data rate under INI in the NOMA-based mixed numerology system. However, they ignored the trade-off in NOMA and numerologies between heterogeneous users, where the best numerology for each user may severely degrade system throughput due to the small difference in channel gains between NOMA grouped users.

In this paper, we explore the PRB allocation in a NOMA-based mixed numerology system with the following new challenges. The first one is the *reused locations of PRBs*.[3] To maximize system throughput, it is desired to overlap allocated

[1]The shape of PRB varies according to numerology [1] and will be detailed in Section II. We will use "numerology" and "shape" interchangeably.

[2]We use the length and width to represent the bandwidth of PRB in the frequency domain and time duration of PRB in the time domain, respectively.

[3]In this paper, we logically regard a PRB as multiple virtual RBs (vRBs) in different layers of a radio frame [9]. Since a PRB includes multiple Basic Units (BUs, detailed in Section II) [10], we call that two vRBs are overlapped if they contain some common BUs to form a reused location of PRB.

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

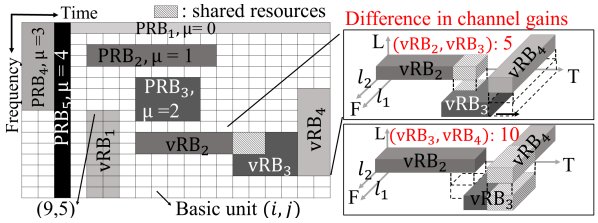


Fig. 1: An example of resource allocation with NR and NOMA.

vRBs with the same numerology [5]; otherwise, INI will occur and degrades throughput [6]. However, the total throughput may increase if the vRBs with different numerologies overlap with each other on some common BUs with a sufficiently large difference in channel gains between users [11]. In Fig. 1, for example, it is inclined to overlap vRB_3 with vRB_4 rather than vRB_2 due to a larger difference in channel gains (i.e., $10 \geq 5$). Hence, it is crucial to decide the reused locations of PRBs since the channel gain of a user may be variant in different channels. The second challenge is the *resource fragmentation*. Due to the different PRB shapes in NR, the allocation sequence of vRBs and their locations in the frame will cause the resource fragmentation. In Fig. 1, if 16×1 PRB_5 is allocated right adjacent to 8×2 PRB_4, the area below PRB_4 will be waste if there are only users requiring 4×4, 2×8, or 1×16 PRBs. Therefore, it is desired to carefully determine the sequence and locations for allocating vRBs with different numerologies. The last challenge is the *heterogeneity of users*. The candidate numerologies of users may be variant (i.e., different shapes of vRBs) due to their QoS requirements. It limits the flexibility in vRB allocation, and a user cannot be served by vRBs with the numerologies other than the candidate ones.

To tackle the above issues, we formulate a new optimization problem, called Multi-Dimensional Resource Allocation Problem (MDRAP), to maximize the weighted sum rate in a radio frame.[4] Different from previous works focusing on NOMA-based single numerology [5], [14] or OFDM-based mixed-numerology systems [15], [16], MDRAP considers the NOMA-based mixed numerology system with different shapes of PRBs reused. We prove MDRAP is NP-hard and design an approximation algorithm, called Flexible Resource Sharing Allocation (FRSA), for MDRAP. FRSA first derives a bounded rectangle, named *zone*, for each user to allocate vRBs, and chooses the shape of vRBs in each *zone* by examining the *Inter-Numerology Relation* (INR), which evaluates the relation between users' candidate numerologies for avoiding INI. Then, FRSA displaces the zones in different layers by examining *Layer Dissimilarity* (LD) and allocates their locations for alleviating INI. LD evaluates the zone dissimilarity in different layers, and a larger LD is likely to cause INI due to the non-perfect zone alignment between different layers. Afterward, FRSA exploits NOMA to improve the total transmission rate by adjusting the locations of vRBs to increase the difference in channel gains between users in different layers. Finally, FRSA carefully examines each remaining space of each layer for allocating more users by examining their weighted rate and available numerology in the space.

The remaining of this paper is organized as follows. Section II describes the system model and MDRAP, and Section III presents the approximation algorithm FRSA. Section IV shows the simulation results, and Section V concludes this paper.

II. Problem Formulation

A. Frame Structure

The radio frame of 5G NR comprises time and frequency domains [17]. In the time domain, the duration of a frame is 10ms and a frame is divided into 10 subframes of always 1ms each for backward compatibility with 4G. Each subframe consists of an integer number of time slots according to the numerology [17]. The basic scheduling unit is a PRB, which is composed of one TTI length in the time domain and 12 consecutive subcarriers in the frequency domain [17]. For agile and efficient resource usage, NR implements a new structure of flexible numerology. The numerology μ defines SCS and TTI, where SCS follows the formula $15 \times 2^{\mu}$ kHz and TTI follows $1 \times 2^{-\mu}$ ms for $\mu \in \mathcal{N} = \{0, 1, \ldots, \mu_{max}\}$ and $\mathcal{N}$ is the set of available numerologies. As shown in Fig. 1, each PRB consists of $2^{\mu_{max}}$ BUs [10], where $\mu_{max} = 4$ denotes the maximum numerology in the system [1]. Each BU occupies the bandwidth of f_{min} Hz and the time slot duration of t_{min} ms. Consequently, the numerology determines the TTI length and the frequency span of a PRB. Specifically, each numerology $\mu \in \mathcal{N}$ of PRB can refer to a specific rectangular shape $2^{\mu} \times 2^{\mu_{max}-\mu}$, where the PRB is with $2^{\mu} \cdot f_{min}$ Hz bandwidth and $2^{\mu_{max}-\mu} \cdot t_{min}$ ms TTI.

Since NOMA allows a PRB can be reused by multiple users, we regard an NR frame logically consisting of multiple layers to represent different layers in the superposition coding scheme of NOMA [9], and a PRB can be regarded as multiple vRBs in different layers. Moreover, each vRB will be mapped to the PRB with the corresponding BUs during transmission. In Fig. 1, for example, vRB_2 in the second layer partially overlaps vRB_3 in the first layer. The overlapped area is the *reused location* of PRBs[3] and will cause additional interference on the data rate as detailed in (1) of the next subsection.

B. System Model and Problem Formulation

We consider a downlink scenario with a single gNodeB having full knowledge of channel state information in 5G-NR [14].[5] Let $\mathbb{U} = \{u_1, \ldots, u_U\}$ be the set of users. Each user includes a minimum data rate demand q_u and a candidate set of numerologies $\mathbb{C}_u \subseteq \mathcal{N}$ [3]. To maintain long-term fairness, a weight ρ_u is the priority of user u set by network operators according to their requirements [12], [13].[4] The radio frame is with the length of $\mathcal{F} = F \cdot f_{min}$ indexed by $\mathbb{F} = \{f_1, f_2, \ldots, f_F\}$ and the width of $\mathcal{T} = T \cdot t_{min}$ indexed by $\mathbb{T} = \{t_1, t_2, \ldots, t_T\}$, where F and T are the numbers of

[4]The weight can be set according to the reciprocal of the average transmission rate to deal with the trade-off between throughput and fairness [12]. It can also support the Service-Level Agreement (SLA) by assigning user weights according to their expenditure on network services [13].

[5]Due to the space constraint, we provide the notation table in [18].

the resource grids in the frequency domain and time domain, respectively. Therefore, a specific BU is denoted by $\{(i,j), i \in \mathbb{F}, j \in \mathbb{T}\}$ with $(1,1)$ located on the top-left of the frame. Since a PRB consists of multiple adjacent BUs, a particular vRB with numerology μ is then mapping to BUs given by $\{(i,j)|f \leq i \leq f+2^{\mu}-1, t \leq j \leq t+2^{\mu_{max}-\mu}-1\}$. As shown in Fig. 1, while vRB_1 contains the BU $(f,t)=(9,5)$ with $\mu=3$, BUs $(9,5)$ to $(16,5)$ and $(9,6)$ to $(16,6)$ are also contained in vRB_1.

Let $\mathbb{B}=\{b_1,b_2,\ldots,b_B\}$ be the set of vRBs,[6] and $\alpha_{b,i,j}$ be an indicator mapping vRB to BU, where $\alpha_{b,i,j}=1$ if vRB b includes BU (i,j); otherwise, $\alpha_{b,i,j}=0$. According to [19], the total number of BUs in one vRB is irrespective of the numerologies and equal to $2^{\mu_{max}}$ (i.e., $\sum_{i=1}^{F}\sum_{j=1}^{T}\alpha_{b,i,j}=2^{\mu_{max}}, \forall b \in \mathbb{B}$). The system has the following constraints. 1) *vRB-user allocation constraint*. Each vRB b can be allocated to only one user u. That is, $\sum_{u\in\mathbb{U}}\beta_{u,b} \leq 1, \forall b \in \mathbb{B}$, where $\beta_{b,u}$ is a binary variable to represent whether vRB b is allocated to user u. 2) *Reusage constraint*. With NOMA, a vRB is allowed to overlap with other vRBs in different frame layers (i.e., they include some common BUs). Let $\mathbb{L}=\{l_1,l_2,\ldots,l_L\}$ be the set of frame layers. Due to the hardware limitation in NOMA-based systems [5], *reusage constraint* indicates that a BU can be reused by at most L users and the maximum number of frame layers is thereby limited by L, i.e., $\sum_{b\in\mathbb{B}}\alpha_{b,i,j} \leq L, \forall i \in \mathbb{F}, j \in \mathbb{T}$. 3) *vRB-layer allocation constraint*. For each layer l, the allocated vRBs are non-overlapped and the total BUs of allocated vRBs cannot exceed the available resource in a frame [17]. Specifically, let B_l be the collection of allocated vRBs in layer l. Each BU can be included in at most one vRB in a layer, i.e., $\sum_{b\in B_l}\alpha_{b,i,j} \leq 1, \forall i \in \mathbb{F}, j \in \mathbb{T}$. The total BUs of allocated vRBs cannot exceed T and F in the time and frequency domains, respectively. That is, $\sum_{j\in\mathbb{T}}\alpha_{b,i,j} \leq T, \forall i \in \mathbb{F}, b \in B_l$ and $\sum_{i\in\mathbb{F}}\alpha_{b,i,j} \leq F, \forall j \in \mathbb{T}, b \in B_l$.

4) *Bandwidth part constraint* [1]. For each user $u \in \mathbb{U}$, the allocated vRBs can only share the same numerology at each time slot. 5) *Robust rate constraint* [20]. The data rate of each vRB allocated to user u is identical to each other (i.e., the vRBs are with the same Modulation and Coding Scheme (MCS)), which is the lowest feasible data rate among all the vRBs allocated to user u. And the total date rate of the allocated vRBs needs to be larger than the minimum data rate demand q_u. Specifically, the total data rate of the vRBs allocated to user u is $R_u=\sum_{b\in\mathbb{B}}\beta_{u,b}\cdot\min_{b\in\mathbb{B}|\beta_{u,b}=1}\{R_{u,b}\} \geq q_u$, where $R_{u,b}$ denotes the data rate of user u on vRB b and can be calculated as [8]

$$R_{u,b}=\mathcal{B}_b log_2(1+\gamma_{u,b}). \tag{1}$$

$\mathcal{B}_b$ is the bandwidth of vRB b and $\gamma_{u,b}=\frac{|h_{u,b}|^2 p_{u,b}}{|h_{u,b}|^2\sum_{u'|h_{u',b'}>h_{u,b}}\lambda_{b,b'}\beta_{u',b'}p_{u',b'}+I_{INI}+\sigma^2}$ is the signal-to-interference-plus-noise ratio (SINR) of user u on vRB b, where $\lambda_{b,b'}=1$ indicates that vRB b overlaps b'; otherwise, $\lambda_{b,b'}=0$. $h_{u,b}$ represents the channel gain of user u on vRB b, $p_{u,b}$ denotes the power allocated to user u on vRB b, and σ^2 is additive white Gaussian noise (AWGN). $I_{INI}=|h_{u,b}|^2\sum_{\hat{b}\in\mathbb{B}|g_{\hat{b}}\neq g_b}\lambda_{b,\hat{b}}p_{u,b}$ is the INI from different numerologies sharing the same radio resources [6], where g_b is the numerology of vRB b. Equipped with the above model, we formulate MDRAP as follows.

Definition 1. Given a set of users $\mathbb{U}=\{u_1,\ldots,u_U\}$ with the corresponding rate demands $\{q_1,\ldots,q_U\}$, and a radio frame with L layers and $F\times T$ BUs, MDRAP is to allocate vRBs to a set of users such that the *vRB-user allocation*, *reusage*, *vRB-layer allocation*, *bandwidth part* and *robust rate* constraints are ensured, and the users' rate demands are met. The objective is to maximize the weighted sum rate $\sum_{u\in\mathbb{U}}\rho_u\cdot R_u$.

Theorem 1. *MDRAP is NP-hard.*

Proof. Due to the space constraint, the detailed proof of NP-hardness is provided in [18].

III. Algorithm

To address MDRAP, an intuitive method is to adopt the measured SINR value [16] to iteratively select the user with the best average channel quality, and allocate vRBs with the maximum numerology chosen from the user's candidate numerologies. The vRBs are scheduled from the first layer of the frame. In each layer, they are scheduled from the left to right and the top to bottom of the frame. However, it ignores the throughput degradation from INI and the vRB allocation sequence, which may result in the resource fragmentation [15].

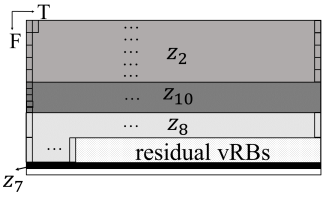
To address the above issues, we design an approximation algorithm FRSA, which includes the following phases: 1) Zone Formation (ZF), 2) Zone Displacement and Allocation (ZDA) and 3) Space Usage Adjustment (SUA). Specifically, ZF first selects the initial numerology for each user according to INR, which examines the relation between users' candidate numerologies to avoid INI. Then, it forms a bounded rectangle, called a *zone*, for allocating vRBs with the identical shape to deal with *resource fragmentation*. Next, ZDA iteratively displaces the zones between different layers by examining LD to alleviate INI and allocates the locations of zones in each layer to minimize the overlap of different shapes of vRBs for dealing with *reused locations of PRBs*. To improve throughput by NOMA, SUA then adjusts the locations of vRBs in each zone for enhancing the difference in channel gains between users in different layers, and serves more users by examining their available numerologies in each remaining space. The overall time complexity of FRSA is $O(U^2L)+O(UFTL)$. Due to the space constraint, the detailed analysis and pseudocode of FRSA are presented in [18]. Each phase is detailed as follows.

1) Zone Formation (ZF): To reduce INI, ZF first selects an initial numerology M_u in $\mathbb{C}_u$ with the maximum INR and constructs a *zone* for each user u. Then, it iteratively selects the zone with the maximum weighted sum rate into the first available layer (i.e., the first one that includes sufficient space (resource) for allocating the zone). Specifically, we first define

[6]The total number of vRBs is the sum of the number of available vRBs for different numerologies in all layers (i.e., $L\cdot\sum_{\mu\in\mathcal{N}}(F-2^{\mu}+1)\cdot(T-2^{\mu_{max}-\mu}+1)$, where L is the maximum number of layers).

user	1	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{C}_u$	{2,3}	{2}	{3}	{0,1}	{3,4}	{1,2,3,4}	{0}	{2,3,4}	{0}	{0}	{4}	{0,1,2}
q_u	223	789	277	831	11	533	53	472	287	382	494	311
ρ_u	0.36	0.71	0.22	0.34	0.7	0.06	0.4	0.64	0.72	0.81	0.1	0.88
S_u	12	40	14	42	1	27	3	24	15	20	25	16
M_u	3	2	3	0	3	3	0	3	0	0	4	0
K_z	8	20	8	21	-	16	2	16	8	10	16	8

(a) Input setting and configuration of users



(b) The zones selected into l_1

l_1

Z	K_z	μ
2	20	2
10	10	0
8	16	3
7	2	0
-	2	-

l_2

Z	K_z	μ
4	21	0
12	8	0
9	8	0
1	8	3
-	5	-

l_3

Z	K_z	μ
3	8	3
11	16	4
6	16	3
-	10	-

(c) The result of ZF

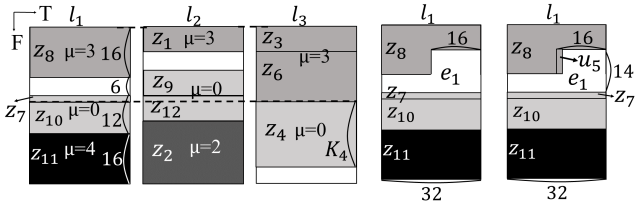
Fig. 2: An illustrative example of ZF.

INR for each numerology as $N_\mu = \sum_{u\in\mathbb{U}} |\mathbb{C}_u \cap \{\mu\}| \cdot |\mathcal{N} - \mathbb{C}_u|, \forall \mu \in \mathcal{N}$, which indicates if μ is in more candidate sets $\mathbb{C}_u$ and $\mathbb{C}_u$ contains fewer numerologies, μ will be assigned a higher priority since it can be utilized by more users to avoid INI. ZF then selects the numerology μ with the maximum INR in $\mathbb{C}_u$, and forms a zone z_u to ensure the number of required vRBs in the worst case (i.e., $S_u = \lceil \frac{q_u}{r_{min}} \rceil$, where r_{min} is the minimum achievable data rate in a vRB with the most robust MCS) can be allocated for each user u. In order to avoid resource fragmentation and INI caused by incomplete alignment [15], ZF categorizes the zones into two sets Z_{merge} and Z_{basic} according to the time duration of required vRBs, where the zones in Z_{basic} are building blocks of zones in Z_{merge} and will be further integrated to reduce resource waste. Specifically, if the time duration of required vRBs of user u is at least $\frac{1}{2}T$ (i.e., $S_u \cdot 2^{\mu_{max}-M_u} \geq \frac{1}{2}T$), ZF adds zone z_u into Z_{merge}; otherwise, Z_{basic}. Each zone $z \in Z_{merge}$ is configured as a rectangle with the length (frequency span) $K_z = \lceil \frac{S_u \cdot 2^{\mu_{max}-M_u}}{T} \rceil \cdot 2^{M_u}$ and width T.[7] For each numerology μ, ZF iteratively selects the zone with μ in Z_{basic} until the total time duration of selected zones is at least $\frac{1}{2}T$ (i.e.,$\sum_{u\in Z_{basic}|M_u} S_u \cdot 2^{\mu_{max}-M_u} \geq \frac{1}{2}T$), and integrates them into a new zone z' by concatenating the zones along the time axis. ZF adds z' into Z_{merge} and configures z' as mentioned above for Z_{merge}. To optimize system throughput, ZF then iteratively selects the zone $z \in Z_{merge}$ with the maximal weighted sum rate into the layer with the smallest index and sufficient space. Since every zone $z \in Z_{merge}$ includes the identical width T, ZF stops when there is no enough residual frequency span (bandwidth) in every layer for allocating zones or all the zones $z \in Z_{merge}$ have been allocated.[8]

Example 1. We assume that $F \times T = 50 \times 32$, $L = 3$, $U = 12$ and set $r_{min} = 19.9$Kbps, and Fig. 2(a) summarizes

[7]To avoid fragmentation in each zone, the residual vRBs in the largest allocated frequency are also included in the zone as shown in Fig. 2(b).

[8]In Theorem 2, we will prove that all layers are allocated at least half of the space of the radio frame to ensure the approximation ratio.



(a) The result of ZDA (b) Space e_1 (c) User u_5

Fig. 3: An illustrative example of ZDA and SUA.

the configuration of each user. For $\mu = 0$, the INR is calculated as $N_0 = 0+0+0+3+0+0+4+0+4+4+0+2 = 17$. With a similar process, we have $(N_0, N_1, N_2, N_3, N_4) = (17, 6, 12, 13, 10)$. For u_1, ZF first selects $M_1 = 3$ because $\mu = 3$ has the maximum INR in $\mathbb{C}_1 = \{2, 3\}$. Then ZF forms z_1 and categorizes it into Z_{merge} since $S_1 = \lceil \frac{223}{19.9} \rceil = 12$ and $12 \cdot 2^{4-3} = 24 \geq 16$. $K_1 = \lceil \frac{24}{32} \rceil \cdot 2^3 = 8$. After examining each user, only z_5 is categorized into Z_{basic} so there is no zone integrated. ZF then sequentially selects z_2, z_{10} to l_1 and z_4 to l_2 according to their weighted rate since there is no sufficient space in l_1 for z_4. Figs. 2(b) and 2(c) present the final result of l_1 and ZF, respectively, where each table in Fig. 2(c) is a layer, and the last row is the residual bandwidth.

2) Zone Displacement and Allocation (ZDA): ZDA consists of two steps ZD and ZA. To alleviate INI for maximizing weighted sum rate, ZD displaces zones between different layers by examining the *dissimilarity* of numerology in different layers, and ZA then allocates the location of zones in each layer accordingly. Specifically, the *dissimilarity* of μ is defined as the average difference in the frequency span between the zones with μ in different layers represented by $\sum_{\hat{l}\in\mathbb{L}|\hat{l}\neq l} \frac{|A_{l,\mu} - A_{\hat{l},\mu}|}{L}, \forall l \in \mathbb{L}$, where $A_{l,\mu}$ is the total frequency span of the zones with μ in layer l. Accordingly, the larger *dissimilarity* is likely to cause INI due to the non-perfect alignment in different layers, and the overall *dissimilarity* of μ is defined as the LD of systems (i.e., $\sum_{\mu\in\mathcal{N}} \sum_{l\in\mathbb{L}} \sum_{\hat{l}\in\mathbb{L}|\hat{l}\neq l} \frac{|A_{l,\mu} - A_{\hat{l},\mu}|}{L}$). For each layer, ZD first chooses the zone with the largest dissimilarity as a candidate to be displaced. Then, it selects two zones z and $\hat{z}$, which have the largest dissimilarity among the chosen zones, and exchanges their layers if 1) there are enough residual bandwidth in each layer (i.e., $K_z \leq K_{\hat{z}} + lr_{\hat{l}}$ and $K_{\hat{z}} \leq K_z + lr_l$, where $lr_l = F - \sum_{z\in Z_l} K_z$ is the residual bandwidth in layer l, and Z_l is the set of the zones allocated in layer l) and 2) LD can decrease to alleviate INI. ZD stops if LD cannot be reduced. Following Example 1, ZD displaces z_2 and z_4 since $\mu = 2$ and $\mu = 0$ have the larger *dissimilarity* in l_1 and l_2, respectively, and LD decreases from 118.67 to 81.33 (we detail the calculation in [18] due to the space constraint). The above process repeats for other zones, and stops when LD cannot decrease. The final result of displacement is $Z_1 = \{z_7, z_8, z_{10}, z_{11}\}$, $Z_2 = \{z_1, z_2, z_9, z_{12}\}$, $Z_3 = \{z_3, z_4, z_6\}$ in Fig. 3(a).

Afterward, ZA selects the layer including the least number of zones as a base layer l^b due to its inflexibility in adjustment and allocates the location of each zone in other layers to align

the zone with the identical numerology in l^b for alleviating INI. Specifically, for each μ in each layer, ZA first concatenates the zones with μ along frequency axis and calculates the total frequency span of μ. For the base layer l^b, ZA then iteratively allocates the concatenated zone from the smallest frequency of l^b by the descending order of the total frequency span. Let f^1_{cz} be the smallest frequency occupied by concatenated zone cz. According to the sequence of utilized μ in l^b from the smallest frequency, ZA iteratively allocates the location of cz with μ for each layer $l \in \mathbb{L} \setminus \{l^b\}$ by aligning f^1_{cz} to $f^1_{cz^b}$ of the concatenated zone cz^b with μ. If there are zones that cannot be allocated due to the limited bandwidth in a frame, ZA sequentially shifts the location of zones from the last zone until all the zones can be allocated in the frame. Fig. 3(a) shows an example. ZA selects l_3 as l^b since it has the least number of zones. For l_1, ZA first aligns z_8 to z_3 in l_3, then it aligns the zone concatenated by (z_{10}, z_7) to z_4. Since z_{11} cannot be allocated into the remaining space of l_1 (i.e., $50-16-8-12=14 \leq 16$), ZA shifts the locations of z_{10} and z_7 until z_{11} can be allocated.

3) Space Usage Adjustment (SUA): SUA maximizes the weighted sum rate by adjusting MCS to release more available spaces and rearranging vRB locations to increase the difference in channel gains between vRBs due to NOMA. Specifically, SUA specifies MCS r_u for each allocated user u according to the channel condition that her vRBs occupy.[9] Then, SUA releases $\lceil \frac{q_u}{r_{min}} \rceil - \lceil \frac{q_u}{r_u} \rceil$ vRBs from the largest frequency to the smallest frequency allocated in z_u because $r_u \geq r_{min}$. Therefore, more spaces are available for adjusting the locations and shapes of vRBs. For each zone z, SUA iteratively selects the vRB b with the smallest difference in channel gains between vRBs in different layers for increasing the data rate of vRB by NOMA. Then, it adjusts b's location to another unoccupied location of the identical frequency in z if the weighted sum rate of the system can increase, where the data rate of each vRB is calculated according to (1). In Fig. 1, for example, we assume the channel gain difference between vRB_2 and vRB_3, and vRB_3 and vRB_4 are 5 and 10, respectively. SUA shifts the location of vRB_3 to neighbor time for increasing the difference in channel gains.

Afterward, SUA allocates more users to maximize the weighted sum rate. For each layer, let $\mathbb{E}_l = \{e_1, e_2, \ldots\}$ denote the set of remaining spaces in l, and CT_i be the maximum time duration in e_i. Since the shape of e_i may be irregular due to the different vRB shapes, SUA iteratively chooses the user that 1) can be allocated in e_i with the numerology $\mu_{e_i} = \max\{\mu_{max} - \lfloor \log_2 CT_i \rfloor, 0\}$ to avoid the rate decreased by the more robust rate of other frequency and 2) includes the maximum weighted rate. Specifically, for each user with $\mu_{e_i} \in \mathbb{C}_u$ and the maximum weighted rate, SUA iteratively allocated vRBs, from the smallest time in e_i, to the frequency that the length of vRB with μ_{e_i} can fit in. If there is no user with $\mu_{e_i} \in \mathbb{C}_u$, SUA increases μ_{e_i} and repeats the above process until there is not sufficient space for allocation. Note that SUA also examines if the weighted sum rate increases in each iteration by calculating the rate of each vRB by (1). Fig. 3(b) shows an example for e_1. We have $CT_1 = 32$ and $\mu_{e_1} = \max\{4 - \lfloor \log_2 32 \rfloor, 0\} = 0$. SUA adjusts μ_{e_1} to 3 since the candidate numerology of unallocated user u_5 is $\mathbb{C}_5 = \{3, 4\}$ in Fig. 2(a). In Fig. 3(c), SUA then allocates u_5 to e_1 at $t = 17$, which is the smallest time that the length of vRB with μ_{e_1} can fit in.

Theorem 2. *FRSA is a $\frac{1}{8}c$-approximation algorithm for MDRAP, where $c = \frac{r_{min}}{r_{max}}$, and r_{min} and r_{max} denote the worst and the best achievable data rate in a vRB, respectively.*

Proof. Due to the space constraint, the detailed proof of the approximation ratio is provided in [18].

IV. Simulation

A. Simulation Setup

We evaluate the performance of FRSA in a mixed numerology NOMA-based system with a single base station (BS) at the center [21]. We consider two user deployment scenarios [12], namely *Cell-edge* and *Random*. In *Cell-edge*, users are distributed near the edge of the coverage area of BS, whereas users are equally-distributed over the whole coverage area of BS in *Random*. We consider five numerologies with (SCS, TTI) respectively being (15kHz, 1ms), (30kHz, 0.5ms), (60kHz, 0.25ms), (120kHz, 0.125ms), and (240kHz, 0.0625ms) [1]. The bandwidth is set to 5 MHz at each frame, and the maximum transmission power of BS is set to 46 dBm [22]. The path loss, shadowing model, and MCS are based on 3GPP specifications [23], [24]. The path loss model follows the macro propagation model for the outdoor urban areas [23], while the shadowing model is log-normal with zero mean and σ^2 variance, where the standard deviation is 8 [24]. The AWGN power spectral density is set to -174 dBm/Hz [22].

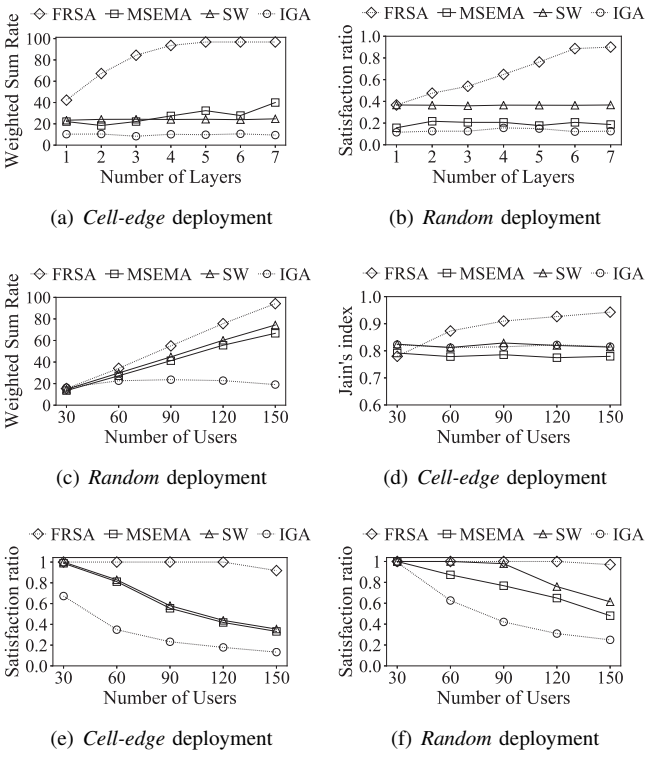
We compare FRSA with 1) Sliding Window-based (SW) scheme [16], 2) Iterative Greedy Algorithm (IGA) [15], and 3) Modified Swap-enabled Matching Algorithm (MSEMA) [25]. SW sequentially allocates PRBs to the user with the best average channel condition, while IGA iteratively schedules the users with the largest utility weight. MSEMA first allocates PRBs to users by examining utility function and then swaps the allocation of grouped users. We change the parameters: 1) number of layers and 2) number of users [25], and evaluate the performance metrics: 1) weighted sum rate, 2) Jain's fairness index [26],[10] and 3) satisfaction ratio, which is the number of allocated users divided by the total number of users. Each result is averaged over 1000 times. Due to the space constraint, we provide more simulations in [18].

B. Simulation results

In Figs. 4(a)(b), we set the number of users to 250. Both the weighted sum rate and satisfaction ratio of FRSA upswings as the number of layers increases because FRSA examines the LD

[9]Note that each zone is formed with MCS r_{min} in ZF, and SUA increases throughput with better MCS to ensure the approximation ratio in Theorem 2.

[10]The index is expressed as $\frac{(\sum_{u=1}^{U} r_u)^2}{U \sum_{u=1}^{U} r_u^2}$ [26], and a higher value means better fairness, where r_u is the average data rate of user u.



(a) *Cell-edge* deployment (b) *Random* deployment

(c) *Random* deployment (d) *Cell-edge* deployment

(e) *Cell-edge* deployment (f) *Random* deployment

Fig. 4: Performance under different parameters.

to reduce INI and exploits NOMA by increasing the difference in channel gains between users in different layers. In contrast, the baselines keep steady because they do not examine different vRB shapes for reusage and will incur more INI. In Figs. 4(c)-(f), we set the number of layers to 2 [5]. FRSA outperforms the other algorithms because FRSA leverages INR to construct *zones* for avoiding resource fragmentation and displaces zones in different layers to decrease LD for alleviating INI. Moreover, SUA carefully adjusts the locations of vRBs to increase the difference in channel gains between users in different layers for maximizing the weighted sum rate to achieve fairness in Figs. 4(c)(d), where the weight prioritizes users for fairness. Comparing Figs. 4(e) with 4(f), *Random* deployment leads to a higher satisfaction ratio since users in the cell-edge need more resources to satisfy their demands due to worse channel quality, and the system thereby serves fewer users in *Cell-edge* deployment.

V. Conclusion

NR and NOMA are two major technologies in 5G to enable flexible and efficient resource allocation. This paper makes the first attempt to explore the resource allocation in a NOMA-based mixed numerology system. We formulate MDRAP, prove the NP-hardness, and design an approximation algorithm FRSA to maximize the weighted sum rate in a radio frame. FRSA leverages INR and LD to alleviate INI and adjusts the location of vRBs in each zone to increase the difference in channel gains between users. Simulation results show that FRSA outperforms state-of-the-art algorithms regarding throughput, fairness, and satisfaction ratio.

References

[1] 3GPP, "NR; physical channels and modulation," 3GPP, Technical specification (TS) 38.211, Sep 2019.

[2] M. Shafi *et al.*, "5G: a tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[3] A. Yazar and H. Arslan, "A flexibility metric and optimization methods for mixed numerologies in 5G and beyond," *IEEE Access*, vol. 6, pp. 3755–3764, 2018.

[4] Y. Liu *et al.*, "Nonorthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[5] L. Dai *et al.*, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 3, pp. 2294–2323, thirdquarter 2018.

[6] J. Choi, B. Kim, K. Lee, and D. Hong, "A transceiver design for spectrum sharing in mixed numerology environments," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2707–2721, May 2019.

[7] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[8] S. McWade, M. F. Flanagan, L. Zhang, and A. Farhang, "Interference and rate analysis of multinumerology NOMA," in *Proc. IEEE ICC*, 2020, pp. 1–6.

[9] H. Zhu, Y. Cao, T. Jiang, and Q. Zhang, "Scalable NOMA multicast for SVC streams in cellular networks," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6339–6352, 2018.

[10] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579–2582, Dec 2018.

[11] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2016.

[12] H. Liao, P. Chen, and W. Chen, "An efficient downlink radio resource allocation with carrier aggregation in LTE-advanced networks," *IEEE Trans. Mob. Comput.*, vol. 13, no. 10, pp. 2229–2239, Oct 2014.

[13] E. Bouillet, D. Mitra, and K. G. Ramakrishnan, "The structure and management of service level agreements in networks," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 4, pp. 691–699, 2002.

[14] F. Fang, J. Cheng, and Z. Ding, "Joint energy efficient subchannel and power optimization for a downlink NOMA heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1351–1364, Feb 2019.

[15] T. T. Nguyen, V. N. Ha, and L. B. Le, "Wireless scheduling for heterogeneous services with mixed numerology in 5G wireless networks," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 410–413, Feb 2020.

[16] W. Sui *et al.*, "Energy-efficient resource allocation with flexible frame structure for heterogeneous services," in *Proc. Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, 2019.

[17] 3GPP, "Release description; release 15," 3GPP, Technical Report (TR) 21.915, Mar 2019.

[18] R.-J. Wang *et al.*, "Resource allocation in 5G with NOMA-based mixed numerology systems (full version)," May 2020. [Online]. Available: http://mnet.cs.nthu.edu.tw/NTHU-TechRep2020.pdf

[19] L. Marijanovic, S. Schwarz, and M. Rupp, "A novel optimization method for resource allocation based on mixed numerology," in *Proc. IEEE ICC*, 2019.

[20] 3GPP, "NR; physical layer procedures for data," 3GPP, Technical specification (TS) 38.214, May 2019.

[21] C. Wang, J. Kuo, D. Yang, and W. Chen, "Surveillance-aware uplink scheduling for cellular networks," *IEEE Trans. Mob. Comput.*, vol. 17, no. 12, pp. 2939–2952, 2018.

[22] M. A. Abu-Rgheff, *5G Physical Layer Technologies*. Wiley, Sep 2019.

[23] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) requirements for LTE pico node B," 3GPP, Technical Report (TR) 36.931, Jan 2016.

[24] ——, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP, Technical report (TR) 38.901, Jun 2018.

[25] B. Liu and M. Peng, "Joint resource block-power allocation for NOMA-enabled fog radio access networks," in *Proc. IEEE ICC*, 2019.

[26] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 1984.