Wireless Scheduling for Heterogeneous Services With Mixed Numerology in 5G Wireless Networks

Ti Ti Nguyen, *Student Member, IEEE*, Vu Nguyen Ha, *Member, IEEE*, and Long Bao Le, *Senior Member, IEEE*

***Abstract*— This letter studies the scheduling problem which determines how time-frequency resources of different numerologies can be allocated to support heterogeneous services in 5G wireless systems. Particularly, this problem aims at scheduling as many users as possible while meeting their required service delay and transmission data. To solve the underlying integer programming (IP) scheduling problem, we first transform it into an equivalent integer linear program (ILP) and then develop two algorithms, namely Resource Partitioning-based Algorithm (RPA) and Iterative Greedy Algorithm (IGA) to acquire efficient resource scheduling solutions. Numerical results show the desirable performance of the proposed algorithms with respect to the optimal solution and their complexity-performance tradeoffs.**

***Index Terms*— 5G NR, wireless scheduling, mixed numerology.**

I. Introduction

THE 5G wireless network is designed to support diverse applications and use cases with different requirements including the enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable and low-latency communication (uRLLC) [1]. Toward this end, the 5G wireless standard adopts the so-called mixed numerology to enable flexible configurations and assignment for different types of physical resource blocks (PRBs) and support different wireless services [2]–[4]. Particularly, while the 4G system supports only one numerology where its PRB has the bandwidth of 180 kHz and time duration of 0.5 ms, the 5G's PRBs can have the bandwidth equal or 2, 4, 8, 16 times of 180 kHz and the time duration equal or 2, 4, 8, 16 times smaller than 0.5 ms. The scheduling problem for resource allocation (RA) of two-dimensional (2D) time-frequency resources has been studied in several recent works [5], [6]. However, the 5G-NR mixed numerology has not been fully studied in [5] while [6] assumes that only two different numerologies exist in the system. Finally, the authors of [7] study the RA problem with mixed numerology for capacity enhancement.

In this paper, we study the scheduling problem for heterogeneous services with mixed numerology which aims to maximize the number of admitted users while meeting service latency and data transmission requirements. Two algorithms, namely Resource Partitioning-based Algorithm (RPA) and Iterative Greedy Algorithm (IGA), are proposed to tackle the underlying problem. The RPA algorithm is developed by partitioning the resources and users into smaller groups, optimizing the RA for each group, then performing resource defragmentation and additional resource assignment for unallocated resources to obtain a final solution. In the IGA algorithm, we iteratively allocate PRBs to users based on an appropriate resource assignment weight to obtain an efficient scheduling solution with low computation complexity. Numerical studies are performed to demonstrate the efficacy of the proposed algorithms.

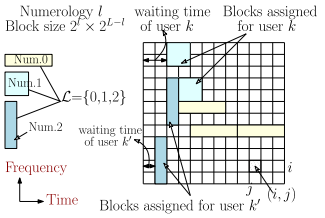


Fig. 1. PRB allocation in 5G wireless networks with mixed numerology.

II. System Model

We consider the 5G system where the available time-frequency resource is divided into resource elements (RE). Each RE occupies the bandwidth of $\Delta^{\mathsf{f}}_{\mathsf{min}}$ (Hz) and the slot duration of $\Delta^{\mathsf{t}}_{\mathsf{min}}$ (seconds). The link/channel conditions for each subcarrier are assumed unchanged during the scheduling time. Moreover, we assume that the 2D RA is performed over each scheduling interval of $T = N^{\mathsf{t}}\Delta^{\mathsf{t}}_{\mathsf{min}}$ (seconds) and the bandwidth of $B^{\mathsf{f}} = M^{\mathsf{f}}\Delta^{\mathsf{f}}_{\mathsf{min}}$ (Hz). Considering the scheduling problem for users where the serving base station supports multiple numerologies as shown in Fig 1. Particularly, the bandwidth of a PRB in numerology l is equal to one half of that in numerology $l+1$ while the time slot duration of a PRB in numerology l is twice of that in numerology $l+1$. The bandwidth of a PRB in numerology l is defined as Δ^{f}_l and the slot duration of a PRB in numerology l is defined as Δ^{t}_l. Then, we have $\Delta^{\mathsf{t}}_l = \Delta^{\mathsf{t}}_{l-1}/2$, $\Delta^{\mathsf{f}}_l = 2\Delta^{\mathsf{f}}_{l-1}$, $\Delta^{\mathsf{t}}_{\mathsf{min}} = \min\{\Delta^{\mathsf{t}}_l, \forall l\}$, and $\Delta^{\mathsf{f}}_{\mathsf{min}} = \min\{\Delta^{\mathsf{f}}_l, \forall l\}$. For convenience, the numerology used by user k is denoted as l_k, the set of all users is denoted as $\mathcal{K}$, the set of numerologies is denoted as $\mathcal{L}$, $l_{\mathsf{max}} = \max\{l \in \mathcal{L}\}$, $l_{\mathsf{min}} = \min\{l \in \mathcal{L}\}$, and $L = l_{\mathsf{max}} - l_{\mathsf{min}}$, and the cardinals of set $\mathcal{L}$ is denoted as $|\mathcal{L}|$.

Each user k requires a data chunk of d^{rq}_k bits be completely transmitted and the total waiting time for its data transmission must not be larger than τ^{max}_k. Note that d^{rq}_k can be the whole data carried by a data flow (e.g., sensing data) or a part of the data of the underlying data flow (e.g., streaming data) in T seconds.[1] We use (i,j) to refer to a particular RE where

Manuscript received October 4, 2019; accepted October 29, 2019. Date of publication November 8, 2019; date of current version February 11, 2020. Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA01. The associate editor coordinating the review of this letter and approving it for publication was A. G. Saavedra. *(Corresponding author: Long Le.)*

T. T. Nguyen and L. B. Le are with the INRS-EMT, Université du Québec, Montréal, QC H5A1K6, Canada (e-mail: titi.nguyen@emt.inrs.ca; long.le@emt.inrs.ca).

V. N. Ha is with the École Polytechnique de Montréal, Montréal, QC H3T1J4, Canada (e-mail: vu.ha-nguyen@polymtl.ca).

Digital Object Identifier 10.1109/LCOMM.2019.2951375

[1]We assume that the d^{rq}_k is known or can be estimated from the QoS requirement and data properties.

its location is given as $f \in [(i-1)\Delta^{\mathsf{f}}_{\mathsf{min}} : i\Delta^{\mathsf{f}}_{\mathsf{min}})$ and $t \in [(j-1)\Delta^{\mathsf{t}}_{\mathsf{min}} : j\Delta^{\mathsf{t}}_{\mathsf{min}})$, for $1 \le i \le M^{\mathsf{f}}$ and $1 \le j \le N^{\mathsf{t}}$.

A. Problem Formulation

PRBs are allocated to users where the numerology is selected in advance by each user. Moreover, each RE is allocated to only one user and associated with one numerology. We represent the mapping for one particular PRB of numerology l to REs in the 2-D frequency-time resource space as follows:

$$\boldsymbol{q}_{l,m,n} = \{(i,j) | m \le i \le m + M_l, n \le j \le n + N_l\}, \quad (1)$$

where $M_l = 2^{l-l_{\mathsf{min}}} - 1$, $N_l = 2^{l_{\mathsf{max}}-l} - 1$. Assuming that the number of PRBs assigned to user k is not larger than C_k to maintain certain fairness among users. Then, we introduce binary variables $x^{k,c}_{i,j}$'s, $y^{k,c}_{m,n}$'s where $y^{k,c}_{m,n} = 1$ if $\boldsymbol{q}_{l_k,m,n}$ corresponds to the c^{th} assigned PRB of user k, and $y^{k,c}_{m,n} = 0$ otherwise; $x^{k,c}_{i,j} = 1$ if RE (i,j) is assigned for user k in its c^{th} PRB, and $x^{k,c}_{i,j} = 0$ otherwise. For $k \in \mathcal{K}$, the ranges of c, i, j, m, and n are $c = 1 : C_k$, $i = 1 : M^{\mathsf{f}}$, $j = 1 : N^{\mathsf{t}}$, $m = 1 : M^{\mathsf{f}} - M_{l_k}$, and $n = 1 : N^{\mathsf{t}} - N_{l_k}$, respectively. We impose the following constraints to ensure non-overlapping RA:

$$\sum_{i'=m:m+M_{l_k}} \sum_{j'=n:n+N_{l_k}} x^{k,c}_{i',j'} \ge 2^L y^{k,c}_{m,n}, \quad \forall k, c, m, n, \quad (2a)$$

$$\sum\nolimits_{k\in\mathcal{K}} \sum\nolimits_{c} x^{k,c}_{i,j} \le 1, \quad \forall i, j, \text{ and } \sum\nolimits_{m} \sum\nolimits_{n} y^{k,c}_{m,n} \le 1, \forall k, c. \quad (2b)$$

Specifically, (2a) implies that the number of REs per PRB remains constant and equal to 2^L. Moreover, (2b) indicates that each RE should belong to only one PRB, and each PRB can be assigned to at most one user. Let $r^k_{m,n}$ denote the transmission rate of user k on PRB $\boldsymbol{q}_{l_k,m,n}$. Then, the total amount of data transmitted by user k during the scheduling interval is $d_k = \Delta^{\mathsf{t}}_{l_k} \sum_{m=1}^{M^{\mathsf{f}}-M_{l_k}} \sum_{n=1}^{N^{\mathsf{t}}-N_{l_k}} \sum_{c=1}^{C_k} r^k_{m,n} y^{k,c}_{m,n}$.

Each user k wants its data chunk to be completely transmitted and the total waiting time be not larger than τ^{max}_k. Let $\tau_{k,0}$ denote the initial waiting time (of the data chunk) of user $k \in \mathcal{K}$ at the beginning of the considered scheduling interval.[2] Then, the total waiting time until the transmission instant of user k can be written as $\tau_k = \tau_{k,0} + \tau_{k,1}$, where $\tau_{k,1}$ is the additional waiting time before user k is served in the scheduling interval, which can be expressed as $\tau_{k,1} = \Delta^{\mathsf{t}}_{\mathsf{min}} \min\{j-1 \mid x^{k,c}_{i,j} = 1, \forall i, j, c\}$.

Our design aims to schedule as many users as possible while meeting their data demand and latency requirements. Recall that user k wishes to transmit a data chunk of d^{rq}_k bits with the total waiting time not larger than τ^{max}_k. To maintain these constraints, we define a function capturing if both constraints are satisfied as $u_k = \mathbb{1}_{d_k - d^{\mathsf{rq}}_k} \mathbb{1}_{\tau^{\mathsf{max}}_k - \tau_k}$, where $\mathbb{1}_x$ stands for the step function, i.e., $\mathbb{1}_x = 1$ if $x \ge 0$, and $\mathbb{1}_x = 0$, otherwise. In fact, if a scheduling solution ensures that the amount of transmitted data and the total waiting time satisfy $d_k \ge d^{\mathsf{rq}}_k$ and $\tau_k \le \tau^{\mathsf{max}}_k$, respectively, we have $u_k = 1$; otherwise, $u_k = 0$. Then, the scheduling problem can be formulated as

$$(\mathcal{P}_1) \max_{\boldsymbol{x},\boldsymbol{y}} \sum\nolimits_{k\in\mathcal{K}} u_k \text{ s.t. } (2a), (2b), \text{ and } \boldsymbol{x}, \boldsymbol{y} \in \{0,1\},$$

where $\boldsymbol{x} = \{x^{k,c}_{i,j} | \forall i, j, k, c\}$ and $\boldsymbol{y} = \{y^{k,c}_{m,n} | \forall k, c, m, n\}$.

B. Problem Transformation

To deal with non-continuous and non-linear functions in problem $(\mathcal{P}_1)$, we first transform this problem into a standard ILP. Specifically, we can equivalently express u_k as

$$u_k \in \{0,1\};\ u_k(\tau^{\mathsf{max}}_k - \tau_k) \ge 0;\ u_k(d_k - d^{\mathsf{rq}}_k) \ge 0. \quad (3)$$

Then, we introduce auxiliary variables $z^{k,c}_{m,n}$'s as $z^{k,c}_{m,n} = u_k y^{k,c}_{m,n}$. Using $z^{k,c}_{m,n}$'s, $(\mathcal{P}_1)$ can be transformed into the ILP form as stated in the following proposition.

Proposition 1: $(\mathcal{P}_1)$ is equivalent to the following problem

$$(\mathcal{P}^{\mathsf{ILP}}_1) \quad \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}\in\{0,1\}} \sum\nolimits_{k\in\mathcal{K}} u_k \text{ s.t. } (2a), (2b),$$

$$\sum\nolimits_{m} \sum\nolimits_{n'=1:N^{\mathsf{rq}}_k} \sum\nolimits_{c} z^{k,c}_{m,n'} - u_k \ge 0, \quad \forall k \in \mathcal{K}, \quad (4a)$$

$$\Delta^{\mathsf{t}}_{l_k} \sum\nolimits_{m} \sum\nolimits_{n} \sum\nolimits_{c} r^k_{m,n} z^{k,c}_{m,n} - u_k d^{\mathsf{rq}}_k \ge 0, \ \forall k \in \mathcal{K}, \quad (4b)$$

$$z^{k,c}_{m,n} \ge u_k + y^{k,c}_{m,n} - 1, \quad z^{k,c}_{m,n} \le \min\{u_k, y^{k,c}_{m,n}\}, \quad \forall k, c, m, n, \quad (4c)$$

where $\boldsymbol{z} = \{z^{k,c}_{m,n}, z^{k,c}_{m,n} | \forall k, c, m, n\}$ and $\boldsymbol{u} = \{u^{\mathsf{r}}_k, u^{\mathsf{d}}_k | \forall k\}$. Here, N^{rq}_k represents the maximum number of time slots (with size of $\Delta^{\mathsf{t}}_{\mathsf{min}}$ seconds) that user k can wait, counting from the beginning of the scheduling interval, to meet its delay constraint which is determined as $N^{\mathsf{rq}}_k = \lfloor (\tau^{\mathsf{max}}_k - \tau_{k_0}) / \Delta^{\mathsf{t}}_{\mathsf{min}} \rfloor$, where $\lfloor . \rfloor$ is the floor operation.

Proof: (4c) with $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u} \in \{0,1\}$ are equivalent to $z^{k,c}_{m,n} = u_k y^{k,c}_{m,n}, \forall k, c, m, n$. As a result, (4a) and (4b) in $(\mathcal{P}_1)$ are equivalent to (3). Hence, $(\mathcal{P}_1)$ is equivalent to $(\mathcal{P}^{\mathsf{ILP}}_1)$. ■

Proposition 2: $(\mathcal{P}^{\mathsf{ILP}}_1)$ is $\mathcal{NP}$-hard.

Proof: If the number of PRBs assigned to each user is known, the remaining problem can be formulated as a well-known *"Cutting Stock Problem"* (CSP), which is strongly NP-hard [8]. Thus, $(\mathcal{P}^{\mathsf{ILP}}_1)$ is more complex than the standard CSP since the numbers of PRBs assigned for different users should also be optimized. Therefore, $(\mathcal{P}^{\mathsf{ILP}}_1)$ must be NP-hard. ■

III. Proposed Algorithms

A. Resource Partitioning Based Algorithm (RPA)

We propose a low-complexity algorithm which solves $(\mathcal{P}^{\mathsf{ILP}}_1)$ by decomposing it into parallel small-scale sub-problems. In fact, if we can maintain the relationship between available resources and users' demands in each sub-problem similar to that in the original problem, then solving small-scale sub-problems can return a solution as good as the one obtained by directly solving the original problem. To realize this idea, we first divide the available bandwidth into M_{B} sub-bands where sub-band m_{B} occupies the spectrum from $(m_{\mathsf{B}} - 1)\lfloor M^{\mathsf{f}}/M_{\mathsf{B}} \rfloor \Delta^{\mathsf{f}}_{\mathsf{min}}$ to $\min\{m_{\mathsf{B}} \lfloor M^{\mathsf{f}}/M_{\mathsf{B}} \rfloor \Delta^{\mathsf{f}}_{\mathsf{min}}, M^{\mathsf{f}} \Delta^{\mathsf{f}}_{\mathsf{min}}\}$. Then, the following three steps are taken in RPA: 1) perform RA on each sub-band,

[2]This initial waiting time is applied to the first chunk of a new data flow when the data flow arrives in the middle of the previous scheduling interval.

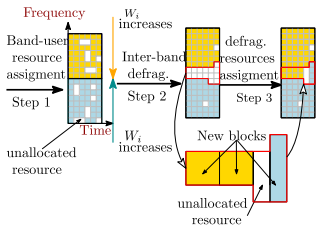


Fig. 2. Illustration of the three steps of RPA algorithm.

2) re-arrange unallocated resources for consecutive sub-bands, and 3) assign these re-arranged resources to support more (unadmitted) users.

Algorithm 1 Resource Partitioning Based Algorithm (RPA)

1: Initialize: Set initial value for M_{B}.
2: Partition resources into M_{B} sub-bands and distribute users into these sub-bands as in **Section III-A**.
3: **Step 1**: Solve $(\mathcal{P}^{\mathsf{ILP}}_{1,m_{\mathsf{B}}})$ to obtain $u_k^{\mathsf{S1}\star}$ for all m_{B}, $k \in \mathcal{K}_{m_{\mathsf{B}}}$.
4: **Step 2**: Solve $(\mathcal{P}_{m_{\mathsf{B}}})$ to create a contiguous region of unallocated resources between two consecutive sub-bands while still satisfying the requirements of admitted users, i.e., users with $u_k^{\mathsf{S1}\star}=1$, $\forall k$.
5: **Step 3**: Solve $(\mathcal{P}_{\mathsf{RPA}})$ to assign unallocated resources to unadmitted users, i.e., users with $u_k^{\mathsf{S1}\star}=0$, $\forall k$.

The key steps of RPA are illustrated in Fig. 2 and it is summarized in **Algorithm** 1. In **Step 1**, we randomly distribute users into sub-bands to make resource demands on different sub-bands similar. Denote the set of users associated with sub-band m_{B} as $\mathcal{K}_{m_{\mathsf{B}}}$, and the set of REs in the frequency dimension as $\mathcal{I}_{m_{\mathsf{B}}}=\{m|m=(m_{\mathsf{B}}-1)\lfloor M^{\mathsf{f}}/M_{\mathsf{B}}\rfloor : \min\{m_{\mathsf{B}}\lfloor M^{\mathsf{f}}/M_{\mathsf{B}}\rfloor, M^{\mathsf{f}}\}\}$. We then solve $(\mathcal{P}_1^{\mathsf{ILP}})$ corresponding to each sub-band m_{B} and the set of users $\mathcal{K}_{m_{\mathsf{B}}}$ to obtain admission decisions, denoted as $u_k^{\mathsf{S1}\star}$'s. The sub-problem for sub-band m_{B} is named $(\mathcal{P}^{\mathsf{ILP}}_{1,m_{\mathsf{B}}})$. In **Step 2**, after finding $u_k^{\mathsf{S1}\star}$'s, we re-arrange the allocated resources so that the unallocated REs from two consecutive sub-bands can be arranged close to one another and they can be combined and mapped into PRBs of certain numerology as defined in (1). The re-arrangement of unallocated REs on sub-band m_{B} can be achieved by solving the following problem:

$$(\mathcal{P}_{m_{\mathsf{B}}}) \quad \min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\{0,1\}} \sum_{k\in\mathcal{K}_{m_{\mathsf{B}}}} \sum_{i\in\mathcal{I}_{m_{\mathsf{B}}}} \sum_{j=1:N^{\mathsf{t}}-N_{l_k}} \sum_{c=1:C_k} W_i x_{i,j}^{k,c}$$
$$\text{s.t. } u_k = u_k^{\mathsf{S1}\star}, (2a), (2b), (4a), (4b),$$
$$\forall k\in\mathcal{K}_{m_{\mathsf{B}}}, i\in\mathcal{I}_{m_{\mathsf{B}}},$$

where $\{W_i\}$ is an increasing series, e.g., $W_i=2^i$ if the sub-band index is odd and $\{W_i\}$ is a decreasing series, e.g., $W_i=2^{-i}$ if the sub-band index is even. It can be verified that after solving $(\mathcal{P}_{m_{\mathsf{B}}})$, all unallocated REs in two consecutive sub-bands will be pushed close to one another to create a contiguous resource region as large as possible. Let $\{\boldsymbol{x}^\star_{\mathcal{P}_{m_{\mathsf{B}}}}, \boldsymbol{y}^\star_{\mathcal{P}_{m_{\mathsf{B}}}}, \boldsymbol{z}^\star_{\mathcal{P}_{m_{\mathsf{B}}}}\}$ denote the optimal solution of $(\mathcal{P}_{m_{\mathsf{B}}})$. In **Step 3**, we assign the unallocated resources, $\Omega=\{(i,j)|\boldsymbol{x}^\star_{\mathcal{P}_{m_{\mathsf{B}}}}=0, \forall m_{\mathsf{B}}\}$, to the set of unadmitted users $\bar{\mathcal{K}}=\{k|u_k^{\mathsf{S1}\star}=0\}$ by solving the following problem $(\mathcal{P}_{\mathsf{RPA}})$:

$$\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}\in\{0,1\}} \sum_{k\in\bar{\mathcal{K}}} u_k$$
$$\text{s.t. } (2a), (2b), (4a), (4b), \forall k\in\bar{\mathcal{K}}, (i,j)\in\Omega.$$

Algorithm 2 Iterative Greedy Algorithm (IGA)

1: Initialize: $d^{\mathsf{rq}}_{k,0}=d^{\mathsf{rq}}_k, c_k=0, \mathcal{W}_{m,n}=1, W=10.$
2: **repeat**
3: Compute $\mathcal{U}_{m,n,k}$ find the largest value of $\mathcal{U}_{m_0,n_0,k_0}$, and perform the corresponding PRB allocation.
4: Update different parameters after the PRB allocation as $c_{k_0}=c_{k_0}+1$, asign $y^{k_0,c_{k_0}}_{m_0,n_0}=1$, update the remaining required data $d^{\mathsf{rq}}_{k,0}=d^{\mathsf{rq}}_{k,0}-r^{k_0}_{m_0,n_0}\Delta^{\mathsf{t}}_{l_k}$, and $\mathcal{W}_{m_0,n}=W, \forall n=1:N^{\mathsf{t}}$.
5: Drop all overlapped PRBs $\boldsymbol{q}_{l_k,m,n}$ to PRB $\boldsymbol{q}_{l_{k_0},m_0,n_0}$
6: **until** $\mathcal{U}_{m,n,k}=0, \forall m,n,k$

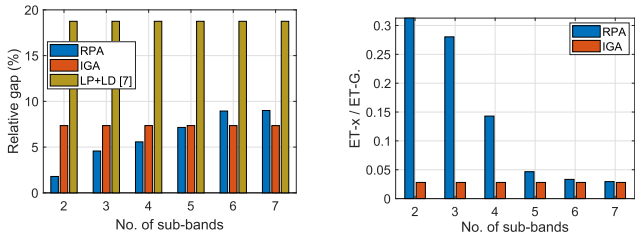
B. Iterative Greedy Algorithm (IGA)

We propose another fast iterative algorithm where we greedily assign resources to users based on an assignment weight which depends on the requirements of the underlying user, the amount of data transmitted and the achieved latency if the underlying PRB is assigned to the user. In each iteration, we calculate the assignment weight for each pair of an available PRB and a user based on which the resource assignment is performed for the PRB-user pair achieving the largest weight. After that, the available PRBs and the weights of all possible PRB-user pairs are updated to prepare for further resource assignment in the next iteration. This process is repeated until there is no more available PRB or unsatisfied user.

We now define the resource assignment weight for a particular PRB-user-resource pair as follows: $\mathcal{U}^k_{m,n}=\frac{r^k_{m,n}\Delta^{\mathsf{t}}_{l_k}}{d^{\mathsf{rq}}_{k,0}}\frac{n}{N^{\mathsf{rq}}_k}\mathbb{1}_{n\in\mathcal{N}_k}\mathcal{W}_{m,n}$ if $c_k\le C_k, d^{\mathsf{rq}}_{k,0}>0$, and $\mathcal{U}^k_{m,n}=0$, otherwise, where $d^{\mathsf{rq}}_{k,0}$ is the remaining required data amount in each iteration, which is equal to d^{rq}_k in the first iteration, c_k is the current total PRBs assigned to user k, and $\mathcal{N}_k$ is the set of REs in the time domain, which is defined as $\mathcal{N}_k=\{n|n\le N^{\mathsf{rq}}_k\}$ if $\sum_{n=1}^{N^{\mathsf{rq}}_k}\sum_{m=1}^{M^{\mathsf{f}}-M_{l_k}}\sum_{c=1}^{C_k} y^{k,c}_{m,n}=0$, and $\mathcal{N}_k=\{n|n\le N^{\mathsf{t}}\}$, otherwise, and $\mathcal{W}_{m,n}$ is a matrix used to mitigate the resource fragmentation in the allocation process, which is updated in each iteration. The unit matrix is initially assigned to $\mathcal{W}_{m,n}$. In particular, $\mathcal{U}^k_{m,n}$ is chosen based on the following criteria: 1) Users with smaller required data chunk receive higher scheduling priorities; 2) If a user is not admitted yet, a PRB at the time location closer to the slot corresponding to the allowed maximum waiting time is more prioritized; 3) Admitted users are allocated resources until their requirements are completely satisfied; and 4) The resource fragmentation is prevented to ease future PRB allocations. The IGA is summarized in **Algorithm 2**. In each iteration of IGA, we need to compute the resource assignment weights $\mathcal{U}_{m,n,k}$ for all available m,n,k, determine the largest $\mathcal{U}_{m_0,n_0,k_0}$ to perform RA, update different parameters, and drop all overlapped PRBs to the assigned block. The worst-case complexity of each iteration is $\mathcal{O}(M^{\mathsf{f}}N^{\mathsf{t}}K)$. Let N^{iter} be the number of iterations, which is upper bounded by $M^{\mathsf{f}}N^{\mathsf{t}}|\mathcal{L}|$, the overall worst-case complexity of IGA is $\mathcal{O}(N^{\mathsf{iter}}M^{\mathsf{f}}N^{\mathsf{t}}|\mathcal{L}|K)$.

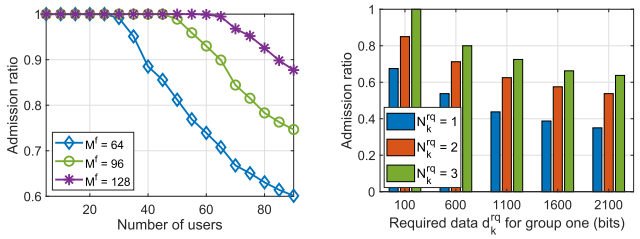
IV. Numerical Results

We consider a wireless system with pedestrian and high moving users in a cell with radius of 500 meters. The channel path-loss β_k (dB) $=128.1+37.6\log_{10}(\gamma_k)$ where γ_k is the



(a) (b)

Fig. 3. Comparison of RPA and IGA with the optimum on the relative gap (a) and the execution time (b).



(a) (b)

Fig. 4. Admission ratio due to IGA for different number of users (a) and for different required data (b).

distance between user k and the BS (in km). For small-scale channel fading, ITU pedestrian-B channel parameters with Doppler shift of 50 Hz and ITU Vehicular-A channel parameters with Doppler shift of 500 Hz are used for pedestrian users and high moving users, respectively. We set three user groups A, B, and C and the numbers of users in these group are $\lfloor K/3 \rfloor$, $\lfloor K/3 \rfloor$ and $K-2\lfloor K/3 \rfloor$, respectively. Specifically, group A adopts numerology 0 corresponding to high moving users with large data demand, group C uses numerology 2 corresponding to high moving users requiring low waiting time, and group B employs numerology 1 corresponding to pedestrian users with average requirements on the transmission data and waiting time. The transmission rate is calculated according to Shannon's capacity where the ratio of transmit power per Hz to noise power density is set equal to 2.8×10^5. The required data chunks d_k^{rq} over the interval T of 1 ms for users in groups A, B, and C are set randomly in $[500-2000]$ (bits), $[500-1000]$ (bits), and $[180-500]$ (bits), respectively, and C_k is set equal to 10. Besides, N_k^{rq} defined in Proposition 1 is set equal to 8 for users in group A and randomly in [3-6] and [1-4] for users in groups B and C, respectively. All numerical results are obtained by averaging the results over 50 random realizations.

We show the relative gap in Fig. 3-(a) which is calculated as $(\sum_k u_k^{\mathsf{G}} - \sum_k u_k^{\mathsf{RPA/IGA}}) \times 100\% / \sum_k u_k^{\mathsf{G}}$ for $M^{\mathsf{f}} = 32$ and $K = 20$ where $u_k^{\mathsf{RPA/IGA/LP+LD[7]}}$ and u_k^{G} represent the objective values for user k obtained by using RPA/IGA/LP+LD [7] and the CVX-Gurobi solver, respectively. Fig. 3-(b) shows the ratio between the average execution time (ET) of the RPA/IGA and that required to solve ($\mathcal{P}_1^{\mathsf{ILP}}$) by the CVX-Gurobi solver (ET-G). The "LP+LD" algorithm proposed in [7] includes two loops: an outer loop to assign PRBs to users based on the utility matrix for all PRB-user pairs, and an inner loop to determine the utility matrix. The figure shows that our proposed algorithms outperform the "LP+LD" algorithm. Besides, as shown in this figure, the relative gap due to RPA increases when the number of sub-bands M_{B} increases. This is because larger M_{B} reduces RA flexibility in each sub-band and thus resource utilization efficiency. However, the execution time of the RPA can be reduced significantly when M_{B} becomes larger. In contrast, IGA always explores good resource-user pairs for efficient resource utilization and IGA is not affected by M_{B}.[3]

[3]By implementing the code on Matlab to solve ($\mathcal{P}_1^{\mathsf{ILP}}$) with Gurobi using a computer equipped with CPU chipset Intel core i7-4790 and 12 GB RAM, average execution time is about 6.39 seconds when $M^{\mathsf{f}} = 32$ and $K = 20$.

Fig. 4-(a) shows the admission ratio achieved by IGA with different values of M^{f} where this ratio is equal to the number of admitted users divided by the total number of users. As can be seen, the admission ratio decreases with the increasing number of requesting users as expected. Moreover, the proposed IGA can admit all users when there are sufficient network resources. We study the interactions between two user groups with different requirements and $M^{\mathsf{f}} = 64$ in Fig. 4-(b). Specifically, group one has varying data transmission demand with maximum waiting time of 1ms while group two requires smaller waiting time compared to group one and each user has a data chunk of $d_k^{\mathsf{rq}} = 250$ bits. In addition, there are 40 users in group one selecting numerology 0 and 40 users in group two using numerology 2 with the same maximum waiting time. For group two, we consider three different values of N_k^{rq}, which are 1, 2, 3. It can be observed that the higher the data transmission demand per user of group one, the lower the admission ratio that can be achieved by IGA. Also, the improvement in the admission ratio when N_k^{rq} increases from 1 to 2 is considerably greater than that when N_k^{rq} increases from 2 to 3. This implies that strict delay requirement may have very negative impact on the system performance.

V. Conclusion

We have proposed two low-complexity algorithms to tackle the scheduling problem for the 5G wireless system supporting heterogeneous services. Numerical results have revealed the desirable performance and complexity for the proposed algorithms.

References

[1] P. Popovski *et al.*, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765– 55779, 2018.

[2] *Physical Channel and Modulation (Release 15)*, document 3GPP-TS-38.211, 2019.

[3] K. I. Pedersen *et al.*, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[4] E. Dahlman *et al.*, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2018.

[5] R. Kassab *et al.*, "Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures," *IEEE Access*, vol. 7, pp. 13035–13049, 2019.

[6] J. Tang *et al.*, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.

[7] L. You *et al.*, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579–2582, Dec. 2018.

[8] C. McDiarmid, "Pattern minimisation in cutting stock problems," *Discrete Appl. Math.*, vol. 98, nos. 1–2, pp. 121–130, Oct. 1999.