2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

Energy-Efficient Resource Allocation with Flexible Frame Structure for Heterogeneous Services

Wenshu Sui, Xiaojing Chen, Shunqing Zhang, Zhiyuan Jiang, and Shugong Xu
Shanghai Institute for Advanced Communication and Data Science
Key laboratory of Specialty Fiber Optics and Optical Access Networks
Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication
Shanghai University, Shanghai 200444, China
E-mail: {wenshusui, jodiechen, shunqing, jiangzhiyuan, shugong}@shu.edu.cn

***Abstract*—The key objective of the fifth generation (5G) wireless technology is to support services with vastly variable requirements, which necessitates the flexible numerology and frame structure for radio resource allocation. In this paper, flexible 2-dimensional resource allocation is investigated to maximize the energy efficiency (EE) for service transmissions with heterogeneous latency requirements. Exploiting the frequency and time diversities of the resource grid, frequency-selective resource allocation, together with an agile "on-off" operation of the power amplifier (PA) are performed by the proposed sliding window-based (SW) algorithm with scalability. Simulation results further demonstrate that the SW algorithm can achieve similar EE performance as the exhaustive method with a substantially lower complexity.**

***Index Terms*—Flexible numerology, flexible frame structure, heterogeneous services, energy efficiency, resource allocation**

I. Introduction

With the increasing diversity of service requirements in the fifth generation (5G) communication systems, the ability to efficiently allocate the limited radio resources for each user adapting to its service requirements becomes extremely desirable. For example, ultra-reliable and low latency communication (URLLC) has been considered as one of the new application scenarios of 5G new radio (NR), enabling mission-critical communications such as vehicular communications, intelligent healthcare, and factory automation [1].

Energy efficiency (EE) is a key metric of 5G mobile communications [2]. EE-maximizing resource allocation strategies for the URLLC scenario have been well studied in extensive works. In [3], an optimal joint power and bandwidth allocation strategy was proposed using a heuristic method, taking into account the transmission and queuing delay, and the reliability constraints. By converting non-convex problems into convex ones, the transmit power allocation, bandwidth configuration and the number of antennas were jointly optimized to maximize the EE under the quality of services (QoS) requirements of URLLC in [4]. A more comprehensive global resource allocation optimization method was further proposed in [5].

To fulfill the low latency requirement in URLLC, one solution is to apply scalable transmission time intervals (TTIs). The flexible numerology and frame structure with configurable subcarrier spacing defined for 5G NR in 3GPP release provide the opportunity to employ a short TTI for resource allocation [6], [7]. A numerology is defined by subcarrier spacing (SCS) and cyclic prefix (CP) overhead. The scalable SCS is defined as $\Delta f = 2^{\mu} \cdot 15$ kHz, $\mu \in \{0, 1, 2, 3, 4\}$ [8]. The flexible frame structure is characterized by the variable number of slots within a frame, depending on the value of μ. A slot comprises 14 orthogonal frequency division multiplexing (OFDM) symbols, leading to a slot length of 1 ms at 15 kHz SCS; a mini-slot comprises 1 to 13 symbols [9]. A TTI can consist of one mini-slot or one slot, or multiple slots if slot aggregation is supported. By using higher numerologies in 5G NR, the OFDM symbol duration, and thus the TTI of a packet decreases, which is beneficial for latency reduction.

The works [3]–[5] did not consider the new changes in the numerology and frame structure of the 5G physical layer. [10] investigated *2-dimensional (2-D)* resource allocation with flexible numerology in frequency domain and variable frame structure in time domain. Compared with fixed numerology and frame structure in 4G communication systems, adopting flexibility in both frequency and time domains showed significant advantages for enhancing the capacity and fulfilling the QoS requirements of URLLC users. In [11], the effects of mixed mathematical-based frame structure on spectral efficiency, computational complexity and signal overhead for resource allocation in 5G systems were explored. By analyzing the relationship between different services and user demands, the authors obtained the most effective number of mixed numerologies using a heuristic method. However, EE-maximizing resource management was not investigated in [10], [11].

With regards to the short-term time-domain techniques considered in 3GPP standard, it is possible to periodically turn off the BS power amplifier (PA) to save energy when there is no downlink traffic [12]. It is indicated in [13] that powering off the PA in time slots that contain signal-free symbols enables up to 47% reduction of the PA operational time. In this paper, we pursue the energy-efficient resource allocation with flexible numerology and frame structure for heterogeneous services. Frequency selective resource allocation, together with an agile "on-off" operation of the PA are investigated to maximize the system EE, while guaranteeing the variable QoS requirements of users. Numerical results show that the proposed strategy has significant EE gain over existing strategies. The main

978-1-7281-2980-8/19/$31.00 ©2019 IEEE
DOI 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00139

contributions of this work can be summarized as follows.

- Adopting flexibility in both frequency and time dimensions, a *2-D* resource allocation problem is formulated to maximize the EE for service transmissions with heterogeneous latency requirements.
- A sliding window-based (SW) algorithm with scalability is proposed, which integrates an agile "on-off" mechanism of the PA operation in accordance to the traffic variation for energy saving. Capturing the frequency-selective and time-varying channel coefficients, the proposed scheme exploits the frequency and time diversities, and obtains the sub-optimal resource allocation with a low complexity.
- Extensive simulations are conducted to demonstrate that the proposed SW algorithm can achieve similar EE performance as the exhaustive method, while substantially reduce the computational time.

The rest of this paper is organized as follows. In Section II, the system model is described. In Section III, the proposed SW algorithm with scalability is developed. Numerical tests are provided in Section IV, followed by concluding remarks in Section V.

II. System Model

Consider the downlink transmission in a single-cell scenario, where a base station (BS) communications with a set $\mathcal{K} := \{1, 2, \ldots, K\}$ of users; see Fig. 1. Heterogeneous services with different data demands and latency requirements arrive at the BS and are to be transmitted to the K users. By exploiting the flexible numerology and frame structure in 5G, proper resource blocks (RBs) with different configurations are assigned to the various services.

As illustrated in Fig. 2, we use basic unit (BU) to refer to the minimum unit of resource in the time-frequency domain [10]. With the time domain of the resource grid indexed by $\mathcal{T} := \{1, 2, \ldots, T\}$, and the frequency domain indexed by $\mathcal{F} := \{1, 2, \ldots, F\}$, a BU in the grid can be uniquely represented by $\{(i, j), i \in \mathcal{T}, j \in \mathcal{F}\}$. There are three types of RBs: i) RB type-1 of shape 4×1; ii) RB type-2 of shape 2×2; and iii) RB type-3 of shape 1×4. Note that each RB type refers to a rectangular shape consisting of four adjacent BUs. Obviously, the incoming services with stringent latency requirement are allocated with RBs type-1, as the short TTI size of RBs type-1 offers the possibility to fulfill the strict latency requirement for users. Neverthless, assigning all services with RBs type-1 is not optimal. Using longer TTIs can exploit larger coding gain to approach the Shannon capacity limit, and also imposes lower control overhead, however, at the cost of increased latency. The services with less stringent latency requirement but higher data rate demand, consequently, are allocated with RBs type-2 or RBs type-3.

Assume that the arrived services have been assigned with the three types of RBs. Let $\mathcal{S} := \{\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3\}$ collect the sets of RBs to be allocated to the resource grid. Here, $\boldsymbol{S_1} := \{\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_M}\}$, $\boldsymbol{S_2} := \{\boldsymbol{B_1}, \boldsymbol{B_2}, \ldots, \boldsymbol{B_N}\}$, and $\boldsymbol{S_3} := \{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_L}\}$ denote the sets of M RBs type-1,

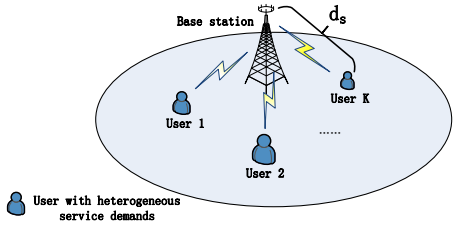

Fig. 1. A base station serves multiple users with heterogeneous service demands.

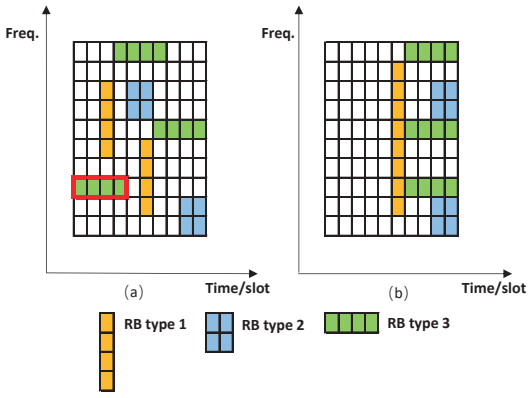

Fig. 2. (a) Random 2-D resource allocation with flexible numerology and frame structure; (b) Optimized resource allocation based on an agile "on-off" operation of the PA.

N RBs type-2 and L RBs type-3, respectively. Recall that a RB consists of four adjacent BUs, and a BU can be represented by $\{(i, j), i \in \mathcal{T}, j \in \mathcal{F}\}$, a particular RB is then given by $\{(i_1, j_1), (i_2, j_2), (i_3, j_3), (i_4, j_4), i_k \in \mathcal{T}_s, j_k \in \mathcal{F}_s, k = 1, 2, 3, 4\}$; $\mathcal{T}_s \subset \mathcal{T}$ and $\mathcal{F}_s \subset \mathcal{F}$ are the location indicators specifying the position of RS s in the resource grid. For instance, the lowest RB type-3 in Fig. 2(a) can be denoted by $\boldsymbol{C}_1 = \{(1, 3), (2, 3), (3, 3), (4, 3)\}$.

Since the RBs must be delivered before their deadlines to fulfill the latency requirement, we have

$$\max\{i_k | i_k \in \mathcal{T}_s\} \leq D_s^{\max}, \tag{1}$$

where the left-hand side of (1) indicates the completion time of delivering RB s to its user, which must not exceed the deadline requirement $D_s^{\max}$.

The achievable data rate of RB s in short blocklength regime can be accurately approximated by [14], [15]

$$C_s = \sum_{i \in \mathcal{T}_s} \left[W_s \cdot \log_2\left(1 + \frac{H_{s,i}^2 P_{s,i}}{\sigma_{s,i}^2}\right) - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon) \right], \tag{2}$$

where W_s is the bandwidth of RB s, $\sigma_{s,i}^2$ is the power of the Additive White Gaussian Noise (AWGN), $P_{s,i}$ is the transmit power allocated to RB s at time i, n is the blocklength, ϵ is the decoding error probability, $Q^{-1}(\cdot)$ is the inverse of the

Gaussian Q-function, and V is the channel dispersion given by [16],

$$V = \frac{H_{s,i}^2 P_{s,i}}{1 + H_{s,i}^2 P_{s,i}}. \tag{3}$$

$H_{s,i}$ is the channel coefficient from the BS to the user requiring RB s at time i, as given by [17]

$$H_{s,i} = \sqrt{L_s(d_s)}\tilde{H}_{s,i}, \tag{4}$$

where $L_s(d_s)$ denotes the path loss from the BS to the user requiring RB s at distance d_s, and $\tilde{H}_{s,i}$ is the small-scale Gaussian distribution coefficient with zero mean.

Consider that the maximum transmit power of the BS $P_{\max}$ is equally distributed over the resource grid per time i [18], then each BU is allocated with power $\frac{P_{\max}}{F}$. The transmit power of RB s at time i is therefore given by

$$P_{s,i} = \frac{P_{\max}}{F}|\mathcal{F}_s|, \tag{5}$$

where $|\mathcal{F}_s|$ is the number of BUs occupied by RB s at time i.

The total power P_i at time i consumed by the BS is [19]:

$$P_i = \begin{cases} \rho + \frac{P_{s,i}}{\eta}, & \text{if } P_{s,i} > 0, \\ P_{\text{sleep}}, & \text{if } P_{s,i} = 0, \end{cases} \tag{6}$$

where ρ is the circuit power consumption when the BS is transmitting (i.e., the PA is in an "on" mode), P_{sleep} is the constant circuit power consumption when the BS is idle (i.e., the PA is in an "off" mode), and $\eta \in (0, 1)$ is the efficiency of the PA at the BS.

The system EE is defined as the ratio of achievable data rate to total power consumption of the BS over time interval T, as given by

$$E(\boldsymbol{I}, \boldsymbol{J}) = \frac{\sum\limits_{s \in \mathcal{S}} C_s}{\sum\limits_{i \in \mathcal{T}} P_i}, \tag{7}$$

where $\boldsymbol{I} = \{\mathcal{T}_s, \forall s\}$ and $\boldsymbol{J} = \{\mathcal{F}_s, \forall s\}$ collect the location indicators of all RBs.

III. RESOURCE ALLOCATION OPTIMIZATION

In this section, we propose a sliding window-based (SW) RB allocation policy to maximize the system EE based on an agile "on-off" operation of the PA, which exploits the frequency and time diversities of the resource grid.

A. Optimization Problem

Our objective is to maximize the EE when serving users by exploiting the frequency and time diversities of the resource grid, and optimally allocating different RBs to the resource grid, without overlapping among the RBs. In addition, the heterogeneous latency requirements of different RBs should be guaranteed. The problem of interest is to solve

$$\max_{\{\boldsymbol{I},\boldsymbol{J}\}} \quad E(\boldsymbol{I}, \boldsymbol{J}) \tag{8a}$$

$$\text{s.t.} \quad \boldsymbol{S_1} \bigcap \boldsymbol{S_2} \bigcap \boldsymbol{S_3} = \varnothing, \tag{8b}$$

$$\max\{i_k | i_k \in \mathcal{T}_s\} \leq D_s^{\max}.$$

Constraint (8b) ensures that no overlapping occurs among the allocated RBs.

B. Proposed RB Allocation Policy

Problem (8) is not convex and is generally NP-hard. One way to solve the problem is resorting to exhaustive search. The exhaustive method is able to find the optimal RB allocation policy by iteratively calculating and comparing the EE for all feasible strategies, and choosing one of the strategy that maximizes the EE as the optimal solution. While exhaustive search inspects all the possibilities of allocating the RBs throughout the resource grid, the computational complexity becomes prohibitively high as the size of the resource grid and the amount of RBs to be allocated becomes large. Specifically, the complexity of the exhaustive method is $\mathcal{O}((F \times T)^{M+N+L})$, where $(F \times T)$ is the size of the resource grid and $(M+N+L)$ is the total amount of RBs waiting to be allocated. The complexity grows exponentially with the amount of RBs, leading to intractability of the exhaustive method in practical implementation.

Nevertheless, the exhaustive method sets a benchmark for a sub-optimal but low-complexity algorithm, which is developed in what follows. Observe first the optimal policies produced by the exhaustive method in Fig. 2(b) and Fig. 3(a). It can be seen that, with the circuit power consumption when the PA is "off" smaller than that when the PA is active, the RBs of different configurations are tended to be gathered and transmitted in a short time range. In other words, as much as time are left to accommodate the sleep mode of the PA. In light of this, we next propose a sliding window-based (SW) algorithm to derive a sub-optimal RB allocation policy with a low complexity.

The proposed SW algorithm is presented in Algorithm 1. It can be concluded from classical information theory that with the same transmit power, transmitting over channels with a better channel state (i.e., a larger channel gain) can provide a larger data throughput, therefore resulting in a larger EE. To this end, we first use a sliding window of size $(F \times window\_width)$ to search for the region with the best average channel state throughout the resource grid, while satisfying the latency requirement of the RBs, as described in Step 5. By sliding the window to obtain a candidate resource poor with the best channel condition, the SW algorithm can benefit from the time diversity of the grid to improve EE.

Choosing a sliding window of size $(F \times window\_width)$ is because a window of this shape can make full use of the "on" time of the PA by scheduling the RBs within a short time range, and thus leave as much time as possible for the PA to "sleep". In addition, as RB type-3 has the longest TTI among the three different RB types, the scheduling of RB type-3 dominates the active time of the BS. By first allocating RB type-3 to the sliding window, and then opportunistically allocating other RB types to the rest BUs in the window according to the channel state, one can, to the utmost extent, shorten the active time of the PA. The width of the sliding window is thus initialized as the TTI size of RB type-3; see Step 3.

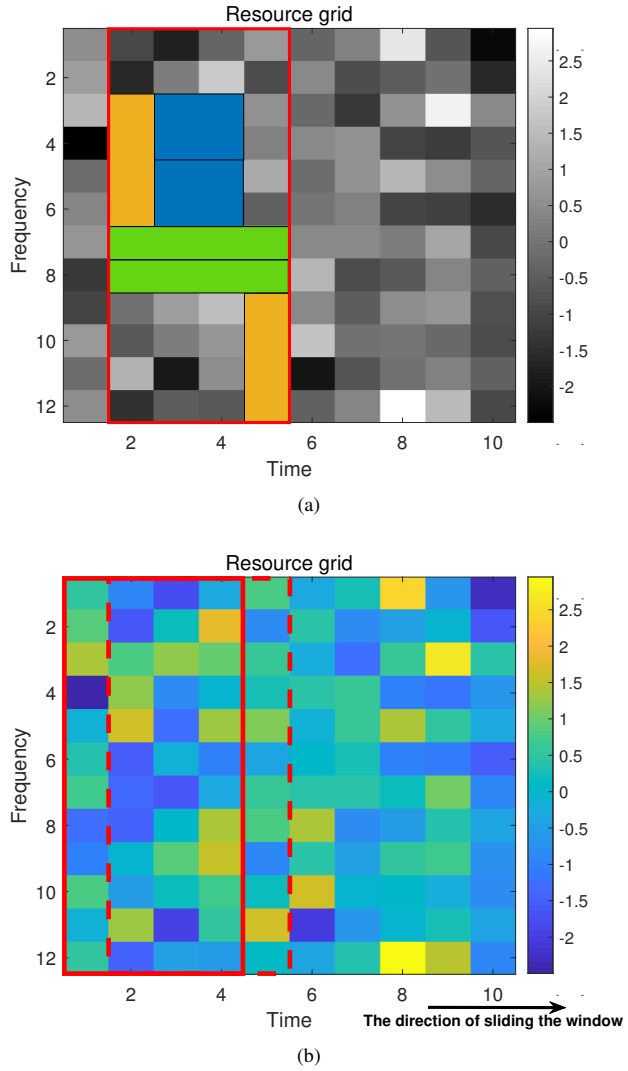

Fig. 3. (a) The optimal resource allocation strategy produced by the exhaustive method, where the delay requirements are $(D_{s_1}^{\max}, D_{s_2}^{\max}, D_{s_2}^{\max}) = (5, 7, 8)$, the number of RBs is $(M, N, L) = (2, 2, 2)$, and the RB type is set as in Fig. 2; (b) An illustration of the SW algorithm. The sliding window selects an available resource pool with the best channel condition.

In Step 6, RBs type-3 are sequentially distributed to the available space in the obtained candidate resource pool of size $(F \times 4)$ according to the channel condition. This step exploits the frequency diversity of the grid. After all RBs type-3 have been allocated, RBs type-2 are first mapped to the unscheduled space in the candidate resource pool with the best channel condition before their deadline. While there is not enough space in the candidate pool to accommodate one more RB type-2, a new sliding window with a width of 2, which is the TTI size of RB type-2, is established and a new candidate resource pool is derived by sliding the window as in Step 5. The procedure continues until all RBs are allocated to the resource grid. The location information of the RBs is returned by $\{\boldsymbol{I}, \boldsymbol{J}\}$, based on which the BS transmits the RBs

Algorithm 1 Proposed sliding window-based algorithm.

1: Initialize the resource grid of size $(F \times T)$; input the frequency-selective and time-varying channel coefficients $\{H_{i,j}^k\}$ from the BS to user k on BU (i, j).
2: Set the number of RBs types-1, 2 and 3 to be allocated as M, N and L, respectively. Set the latency requirements of RBs types-1, 2 and 3 as $D_{s_1}^{\max}$, $D_{s_2}^{\max}$, and $D_{s_3}^{\max}$, respectively.
3: Set the width of the sliding window as $window\_width = 4$.
4: **while** $L > 0$ **do**
5: Allow the sliding window of size $(F \times window\_width)$ to slide along the timeline before RB type-3's deadline in the resource grid, and finalize a candidate resource pool with the best average channel condition, as shown in Fig. 3(b).
6: Search for the unscheduled block (in the same shape as RB type-3) with the best average channel condition inside the candidate resource pool, and accommodate a RB type-3 to the selected block.
7: $L = L - 1$.
8: **end while**
9: **while** $N > 0$ **do**
10: **if** There is not enough space in the window to accommodate a RB type-2 before its deadline **then**
11: Set $window\_width = 2$.
12: Repeat Step 5 with RB type-3 substituted by RB type-2.
13: **end if**
14: Repeat Steps 6 and 7 with RB type-3 substituted by RB type-2, and L substituted by N.
15: **end while**
16: **while** $M > 0$ **do**
17: **if** There is not enough space in the window to accommodate a RB type-1 before its deadline **then**
18: Set $window\_width = 1$.
19: Repeat Step 5 with RB type-3 substituted by RB type-1.
20: **end if**
21: Repeat Steps 6 and 7 with RB type-3 substituted by RB type-1, and L substituted by M.
22: **end while**
23: **return** $\{\boldsymbol{I}, \boldsymbol{J}\}$.
24: Transmit the RBs based on the resultant allocation policy with the transmit power according to (5). Turn off the PA when there is no data to transmit.

to their corresponding users with the transmit power in (5). During the time when there is no data to transmit, the PA is turned off to save energy.

It can be concluded that the complexity of the proposed SW algorithm is lower than $\mathcal{O}(3T + F(M + N + L))$ in the worst case, where $3T$ is complexity to slide the window for a candidate resource pool, and $F(M + N + L)$ is the complexity to search for the unscheduled block with the best

average channel condition inside the candidate resource pool, and accommodate RBs to the selected block. The complexity of the proposed SW algorithm is substantially lower than that of the exhaustive method.

IV. Simulation Results

In this section, simulations are carried out to evaluate the performance of the proposed SW algorithm.

Compared to Long Term Evolution (LTE) that adopts a fixed SCS of 15 kHz and TTI of 1.0 ms, three types of RB, type-1, type-2, and type-3, with SCS being 60 kHz, 30 kHz, and 15 kHz, respectively, are considered in the simulations [20]. The TTIs consisting of 7 OFDM symbols have durations of 0.125 ms, 0.25 ms, and 0.5 ms, respectively, for the three RB types. Note that a TTI of 0.125 ms can meet the worst-case transmission latency for 5G URLLC configuration.

TABLE I
Simulation Parameters

Parameter	Value
Power consumption when PA is off P_{sleep}	4.3 W
Circuit power consumption when PA is on ρ	6.8 W
Noise power $\sigma^2_{s,i}$	-128 dBm
Efficiency of PA η	0.5
Maximum transmit power of BS $P_{\max}$	40 dBm
Bandwidth W_s	20 MHz
Small-scale channel coefficient $\tilde{H}_{s,i}$	Standard normal distribution
Path loss model $10\log(d_s)$	$128 + 37.6\log_{10}(d_s)$
Decoding error probability ϵ	10^{-5}
Size of resource grid $(F \times T)$	(12×10)

The simulation parameters are listed in Table I. The proposed SW algorithm is compared with three alternative algorithms, one is the exhaustive method serving as a benchmark, one is a spectral efficiency (SE)-maximizing scheme (ALG-SE) obtained by exhaustive search, the other is a random resource allocation scheme with the latency requirements guaranteed.

Fig. 4 plots the EE achieved by the proposed SW algorithm, the exhaustive method, the ALG-SE and the random algorithm, as the latency requirements of different RBs become looser (i.e., the deadlines grow longer). The number of the three RB types is set as $M = N = L = 2$. The latency requirements of the three RB types are divided into two cases: i) three RB types have the same latency requirement, and ii) three RB types have different latency requirements. The initial latency requirements are set as $(D^{\max}_{s_1}, D^{\max}_{s_2}, D^{\max}_{s_3}) = (4, 4, 4)$ and $(D^{\max}_{s_1}, D^{\max}_{s_2}, D^{\max}_{s_3}) = (1, 3, 4)$, respectively, and each is increased by one for the next group of trials.

It can be seen from Fig. 4 that the EE of the proposed algorithm and the exhaustive method first increase and then keep unchanged as the deadlines become longer. This is

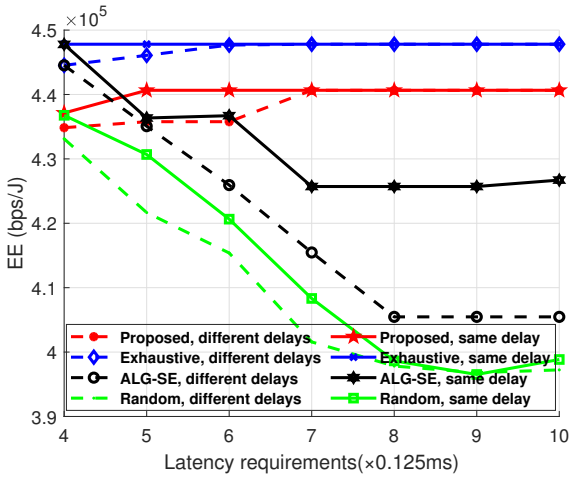

Fig. 4. Comparison of EE among different algorithms, with the growth of deadlines. The number of different RB types is $(M, N, L) = (2, 2, 2)$.

because with longer deadlines, the algorithms can choose the areas with better channel conditions to accommodate the RBs, benefiting more from the time diversity. When the best areas are located before time 7, the increase of deadlines after 7 would not bring extra channel gains to improve the EE, therefore leading to unchanged EE after time 7. Moreover, the optimality gap of the proposed algorithm decreases with the growth of the deadline, and the resultant EE of the proposed algorithm is about 98.4% of its upper bound when the deadline is long. It can also been seen that benefiting less from the time diversity, transmitting RBs with different latency requirements causes EE lost compared to transmitting all RBs with one large deadline.

Interestingly, as the latency requirement becomes looser, the ALG-SE and the random algorithm result in significant reductions in EE. This is because the RBs are widely allocated over the time domain to maximize the SE by the ALG-SE, or just randomly allocated over a wider time range by the random algorithm, which prolongs the "on" time of the PA especially when the latency requirement is loose. The energy consumption is therefore significantly enlarged.

Fig. 5 depicts the EE achieved by the proposed SW algorithm, the exhaustive method and the random algorithm, as the total number of RBs grows. The number of each RB type is set to be the same. We can see that the EE of the two algorithms increase as the number of RBs grows. The increase of the number of RBs to be allocated brings a bigger rise on the system capacity than the energy consumption, which in turn, results in the gradual increase of the EE. However, the increase of EE is limited. Once the number of RBs exceeds a threshold (depending on the size of the resource grid), the RBs cannot be efficiently allocated to the resource grid, but incurs packet loss.

It can also be observed that the EE achieved by the proposed algorithm is very close to that of the optimal exhaustive method when the deadlines are the same, especially with a

TABLE II
CPU RUNTIME COMPARISON

Size of resource grid	48	60	72	84	96	108	120
Exhaustive	47.82 s	1.37×10^3 s	1.28×10^4 s	7.05×10^4 s	$>$ 72 hours	$>$ 7 days	$>$ 10 days
Proposed	0.0374 s	0.0403 s	0.0483 s	0.0511 s	0.0552 s	0.0562 s	0.0526 s
Random	0.0218 s	0.0283 s	0.0266 s	0.0252 s	0.0244 s	0.0258 s	0.0247 s

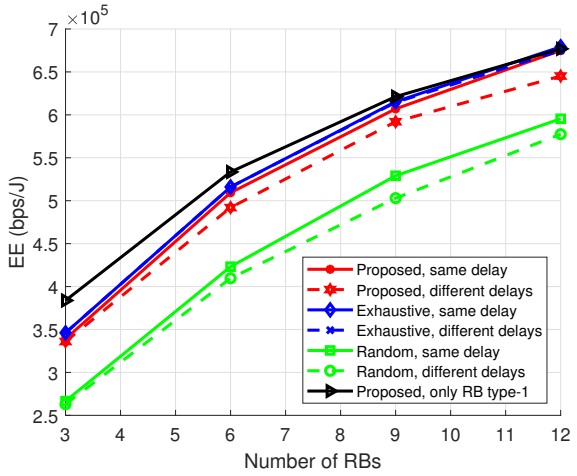

Fig. 5. Comparison of EE among different algorithms, with the growth of total number of RBs. The latency requirements are set as $(D_{s_1}^{\max}, D_{s_2}^{\max}, D_{s_3}^{\max}) = (5, 5, 5)$ (same); and $(D_{s_1}^{\max}, D_{s_2}^{\max}, D_{s_3}^{\max}) = (3, 5, 7)$ (different).

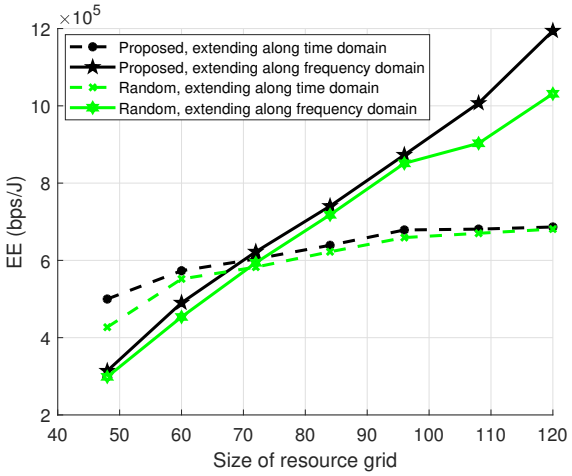

Fig. 6. Comparison of EE among different algorithms, with the growth of resource grid size $(F \times T)$. The size of the resource grid is enlarged either by extending along the time domain with fixed frequency size $(F = 12)$, or along the frequency domain with fixed time length $(T = 6)$.

large number of RBs. Additionally, the proposed algorithm with the same latency requirement always outperforms that with different (or, more stringent) latency requirements, which is because of the better exploitation of the time diversity of the grid, as aforementioned.

The EE achieved by the proposed algorithm with only RBs type-1 to be allocated is also plotted in Fig. 5, which is shown to outperform allocating different RB types. Uniform RBs type-1 can be gathered in a shorter time range, and therefore save as much "on" mode circuit power consumption as possible to improve the EE.

Table II lists the CPU runtime for running the exhaustive method, the proposed method and the random algorithm, where the number of RBs is $(M, N, L) = (2, 2, 2)$. As the size of the resource grid increases, the CPU runtime of the exhaustive method increases exponentially and becomes prohibitively high. The CPU runtime of the random algorithm is the smallest and is not influenced by the change of the grid size. The CPU runtime of the proposed algorithm, however, grows slightly with the increase of the grid size, but is substantially smaller and acceptable compared to the exhaustive method. It can then be concluded that the proposed SW algorithm can achieve similar EE performance as the exhaustive method, while substantially reduce the computational time.

In Fig. 6, the effect of the grid size on EE is explored. The number of RBs is set to be proportional to the grid size. It is shown in Fig. 6 that the EE increases more rapidly when the grid size is enlarged by extending along the frequency domain, and it exceeds the EE obtained by increasing the grid size over the time domain after size 72. It implies that exploiting the frequency diversity leads to a better EE gain than exploiting the time diversity, as the optimal strategy tends to arrange the RBs in a short time range, limiting the EE gain from the time diversity. The scalability of the proposed algorithm is also demonstrated by Fig. 6. The EE performance of the proposed algorithm is always better than that of the random algorithm, as the grid size grows.

V. Conclusion

Considering flexible numerology and frame structure for radio resource allocation in 5G, a flexible 2-D resource allocation was investigated to maximize the EE for heterogeneous services. Exploiting the frequency and time diversities of the resource grid, a sliding window-based algorithm was proposed based on an agile "on-off" operation of the PA. Simulation results show that the proposed algorithm can achieve similar EE performance as the exhaustive method with a substantially lower complexity. Guided by this work, promising future directions include consideration of allocating RBs to services, modeling more practical communication scenario, and developing on-line schemes with a low complexity.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (NSFC) Grants under No. 61701293 and No. 61871262, the National Science and Technology Major Project (Grant No. 2018ZX03001009), the National Key Research and Development Program of China (Grant No. 2017YEF0121400), the Huawei Innovation Research Program (HIRP), and research funds from Shanghai Institute for Advanced Communication and Data Science (SICS).

REFERENCES

[1] S. Zhang, X. Xu, Y. Wu, and L. Lu, "5G: Towards energy-efficient, low-latency and high-reliable communications networks," in *2014 IEEE international conference on communication systems*. IEEE, 2014, pp. 197–201.

[2] A. Mukherjee, "Energy efficiency and delay in 5G ultra-reliable low-latency communications system architectures," *IEEE Network*, vol. 32, no. 2, pp. 55–61, 2018.

[3] C. She and C. Yang, "Energy efficient design for tactile internet," in *2016 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2016, pp. 1–6.

[4] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[5] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2019.

[6] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, 2016.

[7] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio Access Technology; NR; Physical Layer Aspects*, 3GPP, Jan. 2019, ts 38.802.

[8] T. Wirth, M. Mehlhose, J. Pilz, B. Holfeld, and D. Wieruch, "5G new radio and ultra low latency applications: A PHY implementation perspective," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 1409–1413.

[9] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[10] Y. Lei, L. Qi, P. Nikolaos, and Y. Di, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2579–2582, 2018.

[11] 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.

[12] J. B. Rao and A. O. Fapojuwo, "A survey of energy efficient resource management techniques for multicell cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 154–180, 2014.

[13] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 30–38, 2011.

[14] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, p. 2307, 2010.

[15] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[16] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: an information-theoretic study," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[17] S. Bi, Z. Fang, X. Yuan, and X. Wang, "Joint base station activation and coordinated downlink beamforming for HetNets: Efficient optimal and suboptimal algorithms," *IEEE Transactions on Vehicular Technology*, 2019.

[18] Z. Xu, C. Yang, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient MIMO-OFDMA systems based on switching off RF chains," in *2011 IEEE Vehicular Technology Conference (VTC Fall)*. IEEE, 2011, pp. 1–5.

[19] X. Chen, W. Ni, X. Wang, and Y. Sun, "Optimal quality-of-service scheduling for energy-harvesting powered wireless communications," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3269–3280, 2016.

[20] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); NR; Physical channels and modulation*, 3GPP, Jan. 2019, ts 38.211.