Resource Allocation for Block-Based Multi-Carrier Systems Considering QoS Requirements

Arturo González*, Sebastian Kühlmorgen*, Andreas Festag†, Gerhard Fettweis*
*Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany
{arturo.gonzalez,sebastian.kuehlmorgen,gerhard.fettweis}@tu-dresden.de
†Fraunhofer Institute for Transportation and Infrastructure Systems IVI, Dresden, Germany
andreas.festag@ivi.fraunhofer.de

***Abstract*—Future 5G and beyond mobile networks target at services with a high degree of heterogeneity in terms of their communication requirements. To meet these requirements, different PHY numerologies would provide a better performance; still, all services must be served by a single network technology. Generalized Frequency Division Multiplexing (GFDM) is a good candidate for PHY virtualization where the dimensions of the data block can be dynamically configured in time and frequency. Allocating these blocks in a common spectrum every scheduling period leads to a "packing" problem, in which the QoS demands of the data flows need to be acknowledged. In this paper we consider the optimization of the data block allocation as a Knapsack problem. We incorporate the flows' QoS demands by means of utility theory, where utility functions provide a metric of urgency for a flow to be scheduled and the data block to be allocated. For the resulting two-dimensional geometric Knapsack problem we propose a heuristic solution, assess different design options and evaluate the performance in terms of data rate and queuing delay.**

***Index Terms*—5G networks, GFDM, radio resource allocation, 2-D geometric Knapsack packing problem, heterogeneous traffic, QoS, heuristics**

I. Introduction

The design of the 5th generation of mobile networks and beyond (5GB) is strongly driven by demands from industry automation, vehicular communication, smart grid connectivity, and the Tactile Internet, commonly referred as verticals. Their requirements diverge from the traditional voice and broadband services and usually fall into the broad spectrum of machine-type communications. The full deployment of these services will sustainably increase the quality-of-life of people, improve safety as well maximize the efficiency in production and transport. It is expected that 26 % of the worldwide mobile data traffic in 2020 will originate from machine-type communication links [1].

To cope with the diverse communication requirements and channel conditions, several PHY numerologies have been proposed (e.g., [2]), encompassing variable subcarrier bandwidths, symbol duration, frame sizes, cyclic prefixes, modulation and coding, and other parameters. A communication system using different co-existing numerologies need to allocate data blocks of heterogeneous time-frequency dimensions in a common spectrum. Such a system poses a key question: *How to allocate radio resources to heterogeneous data blocks within a common spectrum in an efficient manner such that the data flows' QoS requirements are met?* A straightforward solution divides the spectrum (semi-)statically among different QoS categories, classifies data flows and maps them to the spectrum partitions [3]. However, due to the sporadic characteristics of data traffic, data flows may appear or terminate at any time, rendering the (semi-)static partition approach as inadequate. Therefore, a dynamic radio resource allocation for the different services should be favored.

The problem of dynamic allocation can be modeled as a 2-D packing problem. To incorporate the QoS demands of different services from an upper protocol layer perspective, every block is assigned a utility value representing the performance of its governing flow with respect to the flow's QoS demands. We study the allocation of blocks in the system's radio spectrum and consider the radio resource allocation (RRA) problem as a two-dimensional geometric Knapsack (2D-GK) problem [4].

Modeling RRA as 2-D geometrical packing problem was explored for WiMAX systems, e.g., in [5], [6] as strip- and bin-packing. Their approach partitions rectangular regions in the common spectrum with a single PHY numerology such that these partitions fit the requested data. In contrast, we consider blocks of mixed PHY numerologies and search for an arrangement of defined partitions. Moreover, we study the scheduling and packing of blocks from data flows as a joint problem, unlike the two-stage solution in [5]. Since our system assumptions are fundamentally different from the work in [5], [6], the corresponding algorithms cannot be applied without major modifications to our system. Hence, we do not consider them further nor compare our work to them.

The concepts for RRA presented in this work are applicable for block-based multi-carrier systems. In this paper, we specifically consider Generalized Frequency Division Multiplexing (GFDM) [7], which can be regarded as a generalization of multicarrier modulation, including OFDM. In contrast to OFDM, GFDM allows generating blocks with modulated symbols of variable dimensions regardless of the subcarrier spacing. Due to this flexibility, GFDM can be regarded as a framework for configurable software-defined waveforms [8], where the block dimensions can be tailored to meet the varying QoS requirements of 5GB services. The blocks are transmitted by virtual GFDM transceivers with the configurations and allocations determined by the RRA.

Relying on a radio resource management framework for the GFDM downlink under QoS demands, this paper presents

the design and evaluation of a RRA algorithm for multiuser scheduling with guaranteed QoS demands in terms of data rate and queuing delay. Given the NP-hard complexity of the 2D-GK problem, we propose a heuristic bounded by $O(c \cdot n_b)$ operations, where c is a constant and n_b the number of blocks to allocate. We show that sorting the GFDM blocks under different criteria leads to distinct system performance. For the superior sorting criterion, our results indicate the heuristic, although suboptimal by definition, can still fulfill the demanded QoS requirements of the considered services. To the best of the authors' knowledge, this paper is the first to investigate RRA for GFDM using different co-existent PHY numerologies with QoS support.

The remainder of this paper is organized as follows: Section II provides the required technical background on GFDM and 2D-GK. Section III briefly introduces the radio resource management framework, followed by the description of the RRA algorithm in Section IV. Section VI presents the evaluation environment, including scenario, metrics and parameters, and the results of our simulation. Section VI concludes the paper.

II. Technical Background

A. GFDM

GFDM is a flexible multi-carrier modulation scheme, which is based on the modulation of blocks [7]. Each block consists of a number of subsymbols and subcarriers. The subcarriers are filtered with a prototype filter that is circularly shifted in time and frequency domain. The circularity principle allows GFDM to exploit cyclic prefix (CP) and use frequency domain equalization to handle multipath effects in frequency selective channels. The generalized subsymbol structure of GFDM can be used to combat Doppler effects in time selective channels.

The structure of a GFDM block with the terminology used in this paper is illustrated in Fig. 1. A GFDM block $b[n]$ is composed of K subcarriers carrying M subsymbols, whereas a cyclic prefix (CP) of duration T_{CP} can be prepended to the block. Each subsymbol consists of K data symbols d, resulting in an overall block size of $K \times M$ data symbols. The dimensions of the block in terms of the number of subcarriers K and subsymbols M can vary depending on the flow's QoS requirements in terms of delay and data rate. GFDM blocks will be transmitted in the system-specific time-frequency resources given by the downlink scheduling period τ and the system bandwidth B_{sys} (see Fig. 3 in Section III).

The generation of parallel waveforms with different numerologies is foreseen through software instances of GFDM transceivers (i.e., virtual transceivers) configured with the corresponding waveform parameters. While all virtual transceivers have the same sampling rate, the duration and bandwidth of their operation is determined by the radio resource allocation (RRA) module.

B. Knapsack problem

The Knapsack problem is a classical problem in combinatorial optimization. It is commonly used as model in

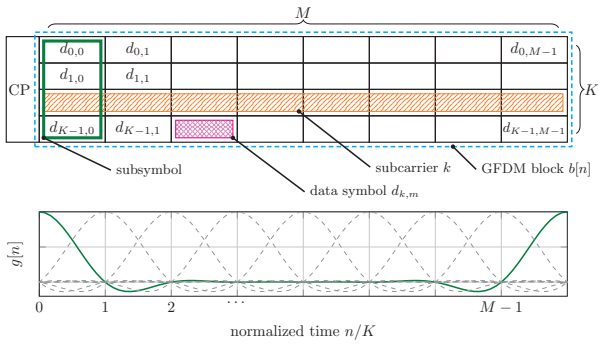

Fig. 1. GFDM block structure (top) and prototype filter (bottom) [7].

engineering, economics and other fields [4]. Intuitively, the base problem considers a set of items, each with a weight and a utility, to be put into a "knapsack", so that the total weight is at most a given limit and the total utility is as large as possible. In the multi-dimensional Knapsack problem, the weight of an item i is given by a n-dimensional vector $\overline{w_i} = (w_1, ..., w_n)$ and the Knapsack has a D-dimensional capacity vector $(W_1, ..., W_D)$. The objective is to maximize the sum of the utilities of the items in the Knapsack so that the sum of weights in each dimension d does not exceed W_d. A more complex variant is the multi-dimensional geometric Knapsack problem (MDGKP), in which geometrical constraints are considered [9]. In the MDGKP, the goal is to find the subset of items that yield the maximum total utility together with their actual packing [10]. We are interested in the 2-D case, by considering a GFDM block as a Knapsack item and the radio spectrum and scheduling period as the 2D Knapsack's capacity.

While the Knapsack problem admits a fully polynomial-time approximation scheme (FPTAS), the two-dimensional geometric version is strongly NP-hard. Moreover, its approximability is not completely understood [10]. Because these problems often arise in practice, it is common to solve them by algorithms that provide workable solutions. An intuitively understandable heuristic is a greedy approach, where instantaneous allocation decisions are taken based on the actual placement situation.

In the context of this work, the task is to choose a set of GFDM blocks and their arrangement into the radio resource common spectrum, such that the sum of the chosen blocks' utility is maximum. In Section IV we will present a heuristic that tackles the 2D-Geometric Knapsack (2D-GK) problem in the context of wireless communication, particularly in GFDM transmission.

III. Framework for Radio Resource Management

Fig. 2 presents the modular resource management framework for GFDM downlink communication. A key concept of the framework is the data flow as an abstraction of services with associated QoS requirements dictated by the services they bear (data plane at the right side of Fig. 2). The assignment of radio resources and transmission configurations to the data flow is determined by the RRA scheme, which exchanges

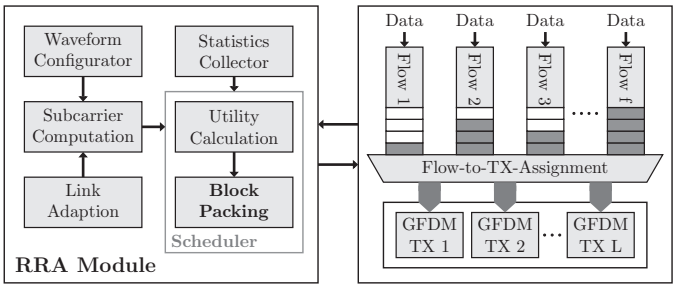

Fig. 2. Resource management framework for GFDM downlink transmission

control information with the data plane. It is assumed that per virtual transmitter, the corresponding receiver is tuned in the transmission bandwidth throughout the transmission duration and that the required information for this is signaled by the base station through a parallel control channel. The framework comprises a set of data queues that hold packets generated by the flows and a function that maps the flows to the corresponding virtual GFDM transceivers.

The RRA consists of several functional components:

The **waveform configurator** matches the QoS requirements of a flow and the link's instantaneous propagation conditions to the best suited waveform configuration. It chooses a GFDM configuration from a pre-defined set and assigns it to a flow f. Each configuration defines the block duration T_b, subcarrier bandwidth B_{sc}, cyclic prefix duration T_{CP} as well as type and parameters of the cyclic filter $\mathbf{g}$ used on the block.

The **link adaptation** selects the modulation and coding scheme (MCS) such that the block error rate can be fulfilled under the given channel conditions.

The **subcarrier computation** determines the number of subcarriers assigned to a block in order to meet the rate demands under the given channel conditions.

The **statistics collector** monitors the behavior of the flows in terms of QoS, including the flow throughput, queue status and waiting time of packets in the queue. It generates statistics (averages, variances, minimum and maximum values and empirical distributions on the run) of the monitored parameters and feeds them to the scheduler.

The **utility calculation** assigns a utility value to the block that represents the observed performance of the flow with respect to the target QoS parameters. Different utility functions can be considered, e.g., data rate, delay, packet loss or a combination thereof.

The **block packing** arranges blocks in the common radio spectrum by solving the 2D-GK problem. It realizes the proposed heuristic that we call *Tightness ranking packing with top cropping* (see Section IV).

IV. Radio Resource Allocation Algorithm

This section presents the Radio Resource Allocation (RRA) problem formulation and the proposed heuristic solution. Every scheduling period τ, the RRA algorithm determines *(i)* the number of subcarriers per block, *(ii)* the utilities of the block and *(iii)* the block arrangement in the common time-frequency plane. Without loss of generality, we assume that only one block is associated to a flow per scheduling period.

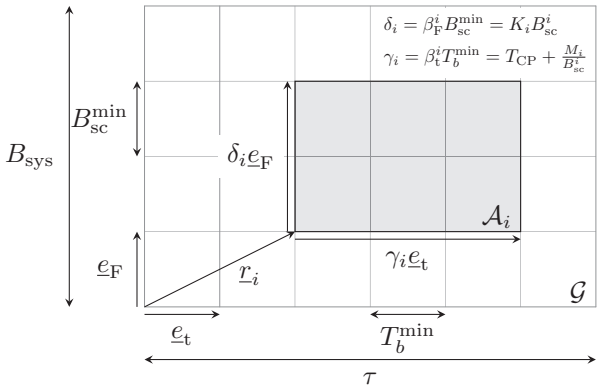

Fig. 3. Arrangement of a GFDM block (in gray) in the time-frequency resource plane.

The number of subcarriers K assigned to a block is a function of the target data rate, the modulation order and the number of subsymbols M in the GFDM block. Depending on the scheduling policy, statistical information of a flow performance can be used to compute K for a block. Every block gets assigned a utility value based on the QoS performance history of the flow they carry. In principle, the utility represents a measure for the flow's underachieved QoS requirements and can be regarded as an indicator of scheduling urgency. The poorer a flow's performance is, the higher the utility value of its block. Clearly, the utility function depends on the scheduling policy and relies on the observations of the statistic collector (Fig. 2). We model the QoS-aware scheduling problem of flows carried by heterogeneous blocks as a 2D-GKP problem: from a set of blocks, each with a utility u_i, find a subset of blocks together with their allocations $\mathcal{A}_i$ in the common time-frequency plane $\mathcal{G}$ such that the sum of utilities is maximum, i.e:

$$\begin{aligned} \boldsymbol{\sigma}^* = & \arg\max_{\boldsymbol{\sigma}\in\boldsymbol{\Omega}_\sigma} \sum_{i=1}^{N} u_i\sigma_i \\ \text{subject to } & (a)\ \bigcup_i \mathcal{A}_i \subseteq \mathcal{G} \\ & (b)\ \mathcal{A}_i \cap \mathcal{A}_j = \emptyset \qquad \forall i \neq j \\ & (c)\ \sigma_i \in \{0,1\} \end{aligned} \tag{1}$$

where

$$\begin{aligned} \mathcal{G} := & \{\alpha_{\text{t}}\underline{e}_{\text{t}} + \alpha_{\text{F}}\underline{e}_{\text{F}} \mid \alpha_{\text{t}}, \alpha_{\text{F}} \in \mathbb{R}_{\geq 0} \ \wedge\ \alpha_{\text{t}} \leq \tau, \alpha_{\text{F}} \leq B_{\text{sys}}\} \\ \mathcal{A}_i := & \{\underline{r}_i + x_i\underline{e}_{\text{t}} + y_i\underline{e}_{\text{F}} \mid x_i, y_i \in \mathbb{R}_{\geq 0} \ \wedge\ x_i \leq \gamma_i, y_i \leq \delta_i \\ & \wedge\ \underline{r}_i = \beta_{\text{t}}^i\underline{e}_{\text{t}} + \beta_{\text{F}}^i\underline{e}_{\text{F}} \ \wedge\ \beta_{\text{t}}^i \ \wedge\ \beta_{\text{F}}^i \in \mathbb{N}_0, \\ & \beta_{\text{t}}^i < \frac{\tau}{T_b^{\min}} \ \wedge\ \beta_{\text{F}}^i < \frac{B_{\text{sys}}}{B_{\text{sc}}^{\min}}\}. \end{aligned} \tag{2}$$

The involved terms are illustrated in Fig. 3.

In (1), the constraint *(a)* indicates that block allocations $\mathcal{A}$ cannot exceed the dimensions of the time-frequency grid $\mathcal{G}$ and the constraint *(b)* restricts overlapping between any two

blocks. The constraint *(c)* expresses that either a block i is selected or not. Finally, the model is defined such that rotation of blocks is not allowed.

Our solution follows the bottom left-justified (BL) packing principle [11], [12]. The objective of these algorithms is to minimize the height of an allocation made up of blocks having equal utilities. We consider packing blocks of different utilities constrained in two dimensions, i.e., in the time-frequency plane. Our approach aims for achieving a tight packing while considering QoS requirements by means of the utilities assigned to the blocks. A tight packing reduces the amount of wasted radio resources and hence increases the system's spectral efficiency. Reduction of wasted resources is accomplished by: *(i)* allowing 'cropping' of blocks in the frequency dimension by an integer number of subcarriers at the cost of a proportional reduction of the block's utility and by *(i)* ranking the possible block placements in terms of *tightness*.

Packing ranking is obtained through linear functions. It is a mechanism for filling-in holes, reduce overall block cropping, and maintaining a uniform packed area that facilitates better subsequent block placements. The ranking function in the frequency dimension assigns a lower ranking when cropping of subcarriers is applied. Packing ranking functions in time and frequency have the same range and are given by

$$\begin{aligned} R_t(\pi, \gamma_b) &= 1 - \frac{\Delta_t}{\tau} \\ \Delta_t &= t_{\text{obs}} - (\pi_t + \gamma_b)\ ,\ 0 < t_{\text{obs}} \leq \tau \end{aligned} \tag{3}$$

$$\begin{aligned} R_F(\pi, \delta_i) &= \begin{cases} \frac{1}{2}(\frac{\Delta_F}{\delta_i} + 1) & -\delta_i \leq \Delta_F < 0 \\ 1 - \frac{\Delta_F}{2B_{\text{sys}}} & 0 \leq \Delta_F \leq B_{\text{sys}} \end{cases} \\ \Delta_F &= B_{\text{obs}} - (\pi_F + \delta_i), \quad 0 < B_{\text{obs}} \leq B_{\text{sys}} \end{aligned} \tag{4}$$

where π is a corner point candidate for the placement of block i. The variables t_{obs} and B_{obs} denote the location of an obstacle in the time and frequency dimensions, respectively.

Eventually, a 2D tightness metric is obtained by combining the individual rankings. For this task, we recognized that the *Cantor pairing* function is well suited since *(i)* it uniquely maps two numbers into one, and *(ii)* for any fixed value of one of the variables, the function is strict monotonically non-decreasing w.r.t. the other variable. Total ranking is calculated as in (5), where R'_F and R'_t are scaled and rounded to the nearest integer versions of (3) and (4), respectively.

$$R_b^{\pi} = \frac{1}{2} \cdot (R'_F + R'_t) \cdot (R'_F + R'_t + 1) + R'_t \tag{5}$$

For the iterative solution, we define the lattice $\mathcal{G}' \subset \mathcal{G}$ with α_{t}, $\alpha_{\text{f}} \in \mathbb{Z}$. A matrix to accommodate the block allocation during the iteration is defined as $\mathbf{A}_{P \times Q}, P = \{1, \ldots \frac{B_{\text{sys}}}{B_{\text{sc}}^{min}}\}, Q = \{1, \ldots \frac{\tau}{T_b^{min}}\}$, where P and Q are the indexes of the midpoints of such lattice $\mathcal{G}'$ in the frequency and time dimensions respectively.[1] A discretization of the blocks' dimensions follows the same approach, however a matrix for their representation is not required for the iteration.

[1]The ratios in the sets P and Q are integers by design.

To avoid scanning the whole time-frequency grid in the allocation process, we save the corner points denoting the origin of regular empty areas that are obtained after a block is placed [12]. Subsequent block placements will only be done at existing points. During the allocation process, possible placements of blocks at these points are ranked as described. A block i is placed at point π for which ranking R_b^{π} is maximum.

As it is common for solutions of Knapsack problems, blocks are sorted in an initial step [4]. For the BL case, sorting the blocks in decreasing order of width generally leads to good results [11], [12]. For our algorithm *Tightness ranking packing with top cropping*, sorting the blocks in decreasing order of the product of the utility and block duration T_b leads to the best results, as it will be shown in Section V.

The algorithm is given in Alg. 1 in pseudo code. It executes in a reduced number of operations bounded by $O(c \cdot n_b)$.[2] Due to space constraints we omit details of how corner points are managed.

Algorithm 1 *Tightness ranking packing with top cropping*

```
A ← 0
for all b do
    R_tot ← 0, π ← {(0,0)}, π_c ← {∅}
    for all π do
        if R'_t(π[j], γ_i > 0) then
            R'_tot ← 0
            if R'_F(π[j], δ_i) > 0 then
                R'_tot = ½(R'_F + R'_t)(R'_F + R'_t + 1) + R'_t
                if R'_tot > R_tot then
                    R_tot ← R'_tot , π_c ← π[j]
                end if
            end if
        end if
    end for
    if R_tot > 0 then
        π ← π \π_c , A ← b
        π ← π ∪ CORNER_POINTS(π_c, δ[i], γ[i])
    end if
end for
```

V. Evaluation

To evaluate the performance of the proposed RRA algorithm by simulation, we have implemented the resource management framework (Fig. 2) in MATLAB. The objective is to estimate the algorithm performance in terms of the system's QoS demands, packing efficiency and stability.The evaluation also assists in the assessment of options for algorithm design, e.g., sorting criteria.

The evaluation scenario assumes a wireless communication system for GFDM downlink transmission relying on different PHY numerologies. We consider a mix of data flows with distinct QoS requirements in terms of data rate and latency.

[2]The constant c is given as $c = n_{\pi}^{\max}(P+Q)$ where $n_{\pi}^{\max}$ is the maximum number of generated corner points, which by observation are much less than the number of blocks n_b. P and Q are the dimensions of the allocation matrix $\mathbf{A}$.

In order to study effects of the RRA algorithm, we choose a scenario in which the system operates at its full capacity. Therefore, the sum of the flows' demanded data rates corresponds to the average system capacity. The average system capacity is a function of those combinations of GFDM blocks from the considered PHY numerologies, which do not leave any empty space after allocation. In this work, we focus on the performance atop the physical layer and are primarily interested in effects of the packing algorithm in isolation from advanced transmission schemes; therefore the simulation model does not consider specific radio conditions nor mechanisms to cope with them. However, extending the model by different modulation and coding schemes with channel feedback is straightforward.

In detail, we study the performance of data flows generated by typical machine-type communication with constant bit rate (CBR) ranging from 16 to 512 kbps and a flow delay budget of 100 ms following periodic data traffic. We have chosen representative system parameters with a bandwidth of 1.92 MHz and BPSK modulation (see Tab. I for the list of simulation parameters). GFDM radio block configurations are randomly assigned to a flow in every scheduling period, i.e., for each flow, the waveform configurator assigns a GFDM block configuration to a flow following a uniform distribution across the set of block durations.

We have devised a set of GFDM block configurations following the GFDM design constraints [7], [8] and basic coexisting rules among them, similar to those presented in [13]. Table II and III show the possible number of subsymbols M of a GFDM block and the corresponding CP duration T_{CP} for a given numerology in terms of block duration T_b and subcarrier bandwidth B_{sc}, respectively. The computation of the number of subcarriers K of a block b aims at achieving the target rate $\rho_{\text{T}}(f)$ of the flow f it bears, i.e.,:

$$K_{b_f} = \min\left(K_{b_f}^{\max},\ \frac{\min\left(D(f),\ \rho_{\text{T}}(f)\tau + \Phi(f)T_{\text{w}}\right)}{M_{b_f}\mu_{b_f}}\right)$$

$$\Phi(f) = \max\left(0,\ \rho_{\text{T}} - \overline{\rho(f)}\right), \quad K_{b_f}^{\max} = \frac{B_{\text{sys}}}{B_{\text{sc}}(b_f)} \qquad (6)$$

where $\overline{\rho(f)}$ is the averaged data rate of flow f computed within a sliding window of length T_w by the statistics collector, M_{b_f} the number of subsymbols of the block b_f assigned to flow f, $D(f)$ are the bits in the queue of f at the computation instant and μ_{b_f} its modulation order in bits/symbol.

The utility assigned to each of the blocks is computed by adding utility components (e.g., data rate and delay) into a total utility value per block. In this paper, we consider guaranteed bit rate (GBR) underachievement[3] and queuing delay as utility components. The utility assigned to a block bearing a flow f is calculated as $u_{\text{Tot}}(f) = u_\rho(f) + u_\delta(f)$ where

$$u_\rho(f) = 0.1^{\frac{\overline{\rho(f)}}{\rho_{\text{T}}(f)}} \quad \text{and} \quad u_\delta(f) = \left(\frac{\delta_{\text{HOL}}(f)}{\delta_{\text{Bud}}(f)}\right)^3 \qquad (7)$$

[3]Data rate below the minimum bit rate requested by an application.

TABLE I
SIMULATION PARAMETERS

Parameter	Values
System bandwidth B_{sys}	1.92 MHz
Sampling rate f_s	30.72 MHz
System capacity	1.680 Msps
Modulation order μ	BPSK (fixed)
Scheduling period τ	1 ms
Data traffic model	Constant bit rate (CBR)
Flow data rate	1×512, 2×256, 4×128, 2×64, 1×16 kbps
Flow delay budget δ_{Bud}	100 ms for all flows
Packet size	1,024, 512, 384, 192, 48 bits
Sorting criteria	UA, UT_b, UB_b, each for ratio and product
Sliding window period T_{w}	500 ms
Scheduling policy	Guaranteed bit rate (GBR) w/ delay constraints
Simulation duration	10,000 scheduling periods

TABLE II
POSSIBLE NUMBER OF SUBSYMBOLS M OF A GFDM BLOCK

T_b (ms)	0.2	0.4	0.6	0.8	1
B_{sc} (kHz)					
7.5	-	-	-	-	7
15	-	-	7	11	13
30	-	9	15	21	27
60	9	21	33	45	57
120	19	43	67	91	115
240	39	87	135	183	231

with $\delta_{\text{HOL}}(f)$ being the head-of-line delay of flow f given by the oldest packet in the queue and $\delta_{\text{Bud}}(f)$ the flow's queuing delay budget. Packets exceeding their flow's queuing delay budget are discarded and accounted as packet drops by the statistics collector.

For the evaluation, we have defined the following metrics:

Data Rate Averaged Moving Mean Error (AMME) quantifies the underachieved data rate by calculating the average difference between the target rate $\rho_{\text{T}}(f)$ and the samples of the moving mean $\overline{\rho(f)}$ whose values are under the target rate. The samples are computed by the statistics collector as a sliding average and collected on every scheduling period over the whole simulation, for all the flows. To obtain a system's outlook of the guaranteed data rate error, we average the AMME of all flows.

Packet Drop Ratio (PDR) indicates the ratio between the number of discarded packets due to aging in the flow's queue

TABLE III
CORRESPONDING CP DURATION

T_b (ms)	0.2	0.4	0.6	0.8	1
T_{CP} (μs)					
7.5	-	-	-	-	66.67
15	-	-	133.34	66.67	133.34
30	-	100	100	100	100
60	50	50	50	50	50
120	41.67	41.67	41.67	41.67	41.67
240	37.5	37.5	37.5	37.5	37.5

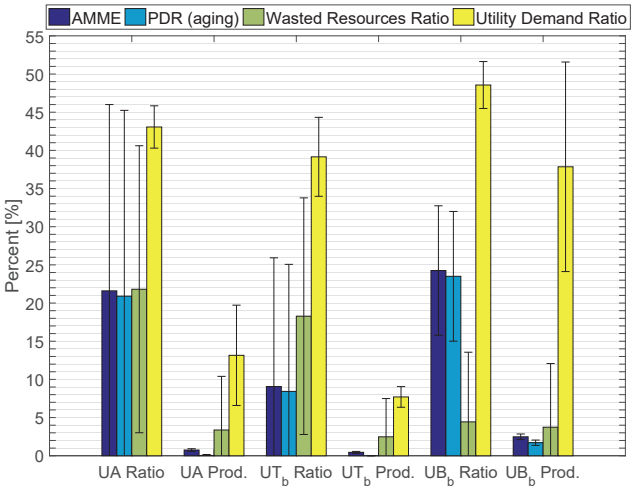

Fig. 4. System performance comparison among six sorting criteria. Sorting in descending order by the product of utility and block duration of the blocks (UT_b Prod.) gives the best performance.

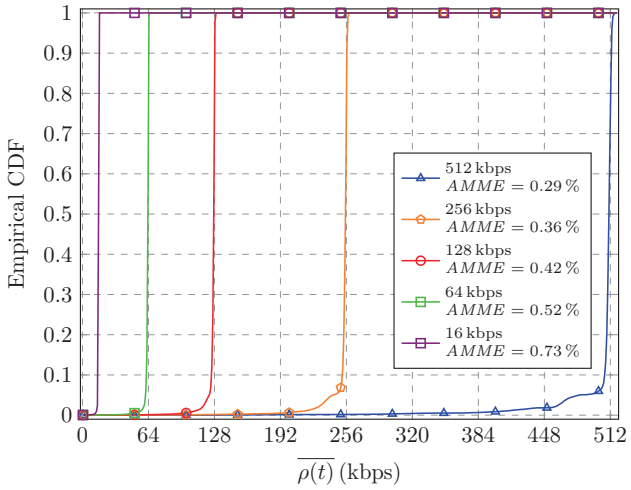

Fig. 5. Moving mean data rate per flow group.

and the number of generated packets of that flow. A system PDR is obtained by averaging the PDRs of the flows.

Wasted Resources Ratio denotes the portion of radio resources that remains unused after a scheduling decision, i.e., the ratio between the non-packed area and the total area of $\mathcal{G}$.

Utility Demand Ratio is obtained for each scheduling period as the ratio between the sum of utilities of all blocks and the maximum utility demand. The latter is computed by the number of utility components times the number of considered flows, considering that the maximum value that each utility component can take is 1, see (7).

In a first step of the evaluation we compare block sorting criteria that combine utility (U) with either block area (A), i.e., $T_b \times B_b$ ($B_b = K \cdot B_{\text{sc}}$), block duration ($T_b$) and block bandwidth B_b – either as a product or a ratio. For example, *UA Ratio* uses the ratio of a block's utility and area as a sorting criterion. Eventually, we yield six sorting criteria.

Fig. 4 compares the system performance of the different block sorting criteria.It can be observed that a greedy approach following the sorting blocks based on decreasing order of utility per dimension unit (i.e., ratios), does not generally lead to a good system performance. We can see that sorting the blocks in a decreasing order of duration and scaled by their respective utilities (UT_b Prod.) leads to the best result. The latter is consistent with the observations reported by [12] and references therein stating that sorting blocks in decreasing order of width, here T_b, leads to better results. The second best system performance is achieved by the sorting criterion of the product of the utility and the area of the block in decreasing order (UA Prod.). For the considered scenario (Tab. I), the scaled utility per block duration T_b shows no packet drops and a minimum AMME (system) of 0.45%. The percent of averaged wasted resources is 2.5 % while the percent of the utility demand ratio is 7.7 %. Other sorting criteria show an increased wasted resources as a consequence of sloppy packing, which itself leads to underachievements in the demanded rates of the system operating at full load. The underachieved rates $\overline{\rho(f)}$ therefore increase the demanded utilities per flow $u_\rho(f)$ and hence the system's utility demand. Moreover, packets at the queues drop because the system cannot deliver the generated packets within the delay budget. These results indicate the significance of tight packing for RRA from a system perspective and the importance of selecting proper sorting criteria.

Fig. 5 shows the empirical cumulative distribution of the samples of the moving mean data rate obtained for all flows. For a better visualization, we have created groups of flows, where flows with the same QoS requirements are jointly collected and plotted. The curves indicate that the demanded rate is achieved on average within a sliding window for all flows. From the curves, it can be appreciated that although some values lie below the target rate, they are not far away from it. This can be corroborated by the obtained values of AMME per flow groups. It is worth noting that the major contribution of error is on those samples obtained at the initial of the simulation, while the system is trying to achieve steady state. Afterwards, the system copes with the demanded rate at full load.

Fig. 6 presents the empirical cumulative distribution of the instantaneous data rate samples collected for all scheduling periods throughout the simulation. The zero values shown for all flows correspond to those scheduling periods where these flows were not scheduled, including the case where no packets were present at their queues due to the relative low data rate of a flow. It can be seen that the higher the flow rate, the more often the flow needs to be scheduled. Clearly, high data rate flows are then the most challenging to deal with. We can confirm from the instantaneous data rate samples that on average the required data rate for all flows is reached. Additionally, it can be observed that fewer than 20 % of instantaneous data rate values for all flows are around the generated bit rate. The latter behavior happens because the system intends to compensate for those data rate values that are considerably below the demanded rate. This is due to the

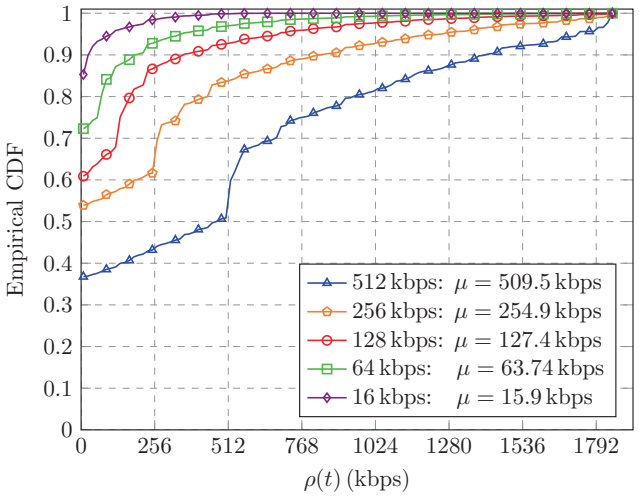

Fig. 6. Instantaneous scheduled data rate per flow group.

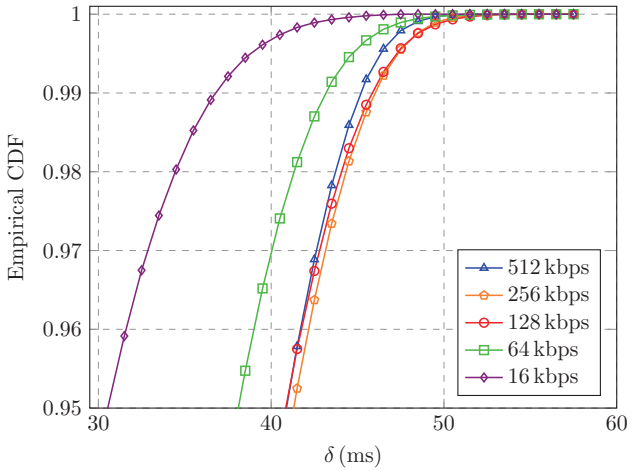

Fig. 7. Queuing delay CDF per flow group from the 95%-tile onwards. No flow group shows any packet loss (PDR = 0% for all flow groups).

calculations of the number of subcarriers K and the data rate utility, which compensate for the average difference between the achieved average rate in a time window and the demanded rate. When the system operates at full load we can observe that with CBR it tends to find equilibrium by generating blocks with several subcarriers (i.e., tall blocks) and taking advantage of idle periods of other flows to schedule them.

Finally, Fig. 7 shows the 95 %-tile of the queuing delay values for all flow groups. All delay values lie way below the considered queuing delay budget of 100 ms with values below of 55 ms. Consequently, no packets are dropped. These results suggest that for the considered CBR traffic more demanding delay budgets can be supported.

VI. Conclusions

In this paper, we presented a QoS-aware RRA framework supporting the allocation of data blocks with heterogeneous and co-existing numerologies in a common spectrum. We considered a GFDM system that facilitates the design of such numerologies and enables determining the dimensions of the independent blocks. We showed that the RRA in such systems can be modeled as a 2D-GK problem and developed the *Tightness ranking packing with top cropping* heuristic as an approach to solve it. The solution aims at reducing the wasted radio resources by means of tight packing. It considers the QoS demands of flows by assigning utilities to the handled blocks based on the flow's performance. Tight packing is achieved by cropping of the blocks in the frequency dimension and by ranking the possible placements of the blocks in terms of tightness. Numerical simulations showed that the presented solution meets the QoS demands of heterogeneous flows in terms of average data rate and queuing delay.

In future work we plan to apply admission control and traffic shaping techniques to guarantee a constant-bit rate to the scheduled instantaneous data rate. Moreover, we plan to customize and evaluate the approach in vehicular communication scenarios with mixed data traffic.

Acknowledgment

This work was supported by the German Science Foundation (DFG) within the priority program CoInCar (SPP 1835).

The authors thank Dr. Dan Zhang from Technische Universität Dresden for valuable discussions.

References

[1] "Visual Networking Index: Forecast and Methodology, 2015-2020," White Paper, Cisco, Jun. 2016. [Online]. Available: http://goo.gl/I8QDy4

[2] A. A. Zaidi *et al.*, "Waveform and Numerology to Support 5G Services and Requirements," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 90–98, Nov. 2016.

[3] G. Wunder *et al.*, "5GNOW: Non-orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Comm. Mag.*, vol. 52, no. 2, pp. 97–105, Feb. 2014.

[4] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*. Bologna: John Wiley and Sons Ltd., 1990.

[5] C. Cicconetti, L. Lenzini, A. Lodi, and S. Martello, "Efficient Two-Dimensional Data Allocation in IEEE 802.16 OFDMA," *IEEE/ACM Trans. Net.*, vol. 22, no. 5, pp. 1645–1658, Oct. 2014.

[6] C. So-In, R. Jain, and A. Abdel-Karim, "eOCSA: An Algorithm for Burst Mapping with Strict QoS Requirements in IEEE 802.16e Mobile WiMAX Networks," in *IEEE 2nd IFIP WD*, Paris, France, Dec. 2009, pp. 1–5.

[7] N. Michailow *et al.*, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Trans. Comm.*, vol. 62, no. 9, pp. 3045–3061, Sep. 2014.

[8] I. Gaspar *et al.*, "GFDM – A Framework for Virtual PHY Services in 5G Networks," *CoRR*, 2015. [Online]. Available: https://arxiv.org/abs/1507.04608

[9] J. Cagan, "The Constrained Geometric Knapsack Problem and its Shape Annealing Solution," Technical Report, Carnegie Mellon University, 1992. [Online]. Available: http://repository.cmu.edu/meche/35/

[10] A. Adamaszek and A. Weise, "A quasi-PTAS for the Two-Dimensional Geometric Knapsack Problem," in *ACM SIAM '2015*, San Diego, CA, USA, Jan. 2015, pp. 1491–1505.

[11] B. Baker, E. Coffman, and R. Rivest, "Orthogonal Packings in Two Dimensions," *SIAM Journal on Computing*, vol. 9, no. 4, pp. 846–855, Nov. 1980.

[12] B. Chazelle, "The Bottom-Left Bin-Packing Heuristic: An Efficient Implementation," *IEEE Trans. Comp.*, vol. C-32, no. 8, pp. 697–707, Aug. 1983.

[13] K. Pedersen, G. Berardineli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases," *IEEE Comm. Mag.*, pp. 53–59, Mar. 2016.