Scalable and QoS-aware Resource Allocation to Heterogeneous Traffic Flows in 5G

Yassine Boujelben

***Abstract*—Networks of new generations are increasingly involved in transporting heterogeneous flows. Indeed, in addition to the usual data and multimedia traffic, the Internet of Things (IoT) smart applications are creating new traffic types and relationships involving billions of active nodes like sensors and actuators. This traffic raises a problem of scale, particularly for resource management and decision-making mechanisms. The present work addresses for the first time the joint problem of mapping heterogeneous flows from multiple users and applications to transport blocks, and then packing these blocks into the rectangular grid of time-frequency resources within a flexible 5G new radio frame. Our solution is based on a quality-of-service-based classification of flows followed by an offline construction of two databases. The first one enumerates all possible configurations of transport blocks, and the second enumerates all possible configurations of frames. Thus, the sole online processing that remains to be done is to find the optimal block configurations that satisfy a given request vector. Hence, the resolution of this complex joint mapping and packing problem is reduced to a simple resolution of a linear problem which consists in finding the best configurations. A thorough numerical study shows that our configuration-based solution can map, within few tens of milliseconds, more than 100 flow connections to transport blocks incurring only 3% of overallocation, and then pack these blocks into the grid leading to an upper bound on the optimality gap as low as 2.8%.**

***Index Terms*—Resource management, 5G NR, URLLC, eMBB, Internet of Things (IoT), 2D Packing problem.**

I. Introduction

The advent of the Internet of Things (IoT) is creating significant challenges for networking in general, and especially for the access networks. The IoT was spurred by the idea of interacting with sensors and actuators in an industrial environment to control production entities remotely, i.e. through wireless access networks and the Internet [1]. This idea was quickly adopted by several other sectors such as agriculture, transportation and healthcare [2]. Then, it was generalized to cover all facets of our daily life, especially with smart applications such as the smart cities, smart houses, smart grids, etc. [3], [4]. With a network of such growth and scope, several challenges have been identified such as architecture, performance, resource management, scalability and security [2]. These challenges have intensified with the development of 5G networks that are expected to cover these new IoT applications simultaneously with the traditional broadband applications [5].

The heterogeneity of flows and services that should be carried simultaneously on the same frame is an important challenge for the design of the 5G networks. Actually, in addition to the conventional enhanced mobile broadband (eMBB) traffic which is increasingly demanding in bandwidth, flows generated by massive machine-type communications (mMTC) have far weaker requirements in terms of resources [6], but some applications, e.g. mission critical (MCC) or ultra-reliable low-latency communications (URLLC), require very strict time and reliability constraints [7]. Thus, the allocation of resources should be done with different levels of granularity, unlike 4G networks where the resource allocation unit is fixed. One of the solutions currently advocated is to have a flexible frame structure [8]. This solution consists in allowing variable transmission time interval (TTI) periods depending on the traffic demand [8]–[13]. With such flexibility, the two-dimensional (2D) frame will be formed by several regions or blocks of different sizes multiplexed in time and frequency. The optimization of the time-frequency resources will depend, among others, on where to place each user's block inside the frame. The effectiveness of this 2D bin packing-like procedure is closely related to the sizes of each block [14], [15]. Consequently, special attention will be paid to the choice of the block sizes.

Moreover, an appropriate Quality-of-Service (QoS) architecture is required in order to provide a specific performance guarantee for each service. The 5G QoS model uses the concept of *QoS flow* in order to classify the IP data flows traveling in the downlink direction [16]. Thus, all data flows associated with the same QoS flow receive the same processing during the forwarding process over the 5G core (5GC) network. At the edge of the next generation radio access network (NG-RAN), the QoS flows are *mapped* to Data Radio Bearers (DRB) which are in turn scheduled and mapped to transport channels by the medium access control (MAC) scheduler of the base station. The binary data in these transport channels are finally allocated 2D blocks of time-frequency resources. This process will be described in more detail in section II-B. Nevertheless, it can already be noted the crucial function of the MAC scheduler which is responsible for the one-to-one mapping between the QoS flows and the radio resources. Unfortunately, this mapping could fail for some QoS flows due to a lack of radio resources, which will force the user plane function (UPF) to discard the downlink data packets associated with the corresponding QoS flows leading to possible QoS level degradation [17]. Therefore, a careful management of

The author is with the École Nationale d'Électronique et des Télécommunications de Sfax (ENET'Com), University of Sfax, Tunisia (e-mail: yassine.boujelben@enetcom.usf.tn), and also with the New Technologies and Telecom Systems Research Unit (NTS'COM).

the radio resources must be undertaken in order to minimize the number of scheduled-but-not-served QoS flows.

In this work, we will consider the joint problem of mapping the QoS flows to transport blocks and packing these blocks in the 2D frame based on the novel objective function of minimizing the number of scheduled-but-not-served QoS flows. In addition to considering the efficiency of scheduling through problem modeling, we also propose an innovative procedure of resolution by maximizing the offline processing so that, the online decision will be limited to a simple optimization of a linear problem whose resolution is performed in a few tens of milliseconds. Acutally, our solution is based on a prior definition of the flow types and block classes, then on the construction of two databases: one for the block configurations and the other for the different packing arrangements inside the frame. Thus, when the scheduler provides a list of flows to be sent on the next frame, the only processing to be carried out in real-time is to choose from the configuration database, those which allow these flows to be transported at a lower cost, then to seek the appropriate packing in the frame database. Consequently, the resolution is reduced of this complex problem of resource allocation and block packing to a simple resolution of a linear problem for the choice of the configurations. To the best of our knowledge, this work is the first to have considered the problem of dynamic allocation of resources for heterogeneous flows in 5G in this innovative form both at the formulation and resolution levels.

A. Related works

Resource allocation for heterogeneous traffic in 5G is of increasing importance. Several studies have approached this problem from different angles. In a recent survey [18], many optimization metrics have been outlined and the general observation made by the authors underlines the high computational complexity of the majority of the proposals. The coexistence of eMBB and URLLC services has generated the most interest in previous work dealing with heterogeneous flows. In [7], three solutions for the coexistence problem have been described. First, dynamic scheduling offers the best cost and resource efficiency, but requires the careful handling of the very different traffic profiles and QoS requirements. Second, the reservation technique for URLLC which would incur waste of radio resources. Finally, the puncturing/superposition technique, which has inspired the greatest interest, allows URLLC flows to overwrite the eMBB blocks, which should push the affected flows to find the necessary means to report and correct any errors. In this section, a different classification approach will be used based on the number of dimensions used for resource optimization. Therefore, the 1D and the 2D categories are defined.

The one-dimensional (1D) category emphasizes the coexistence of eMBB and URLLC/MCC services and considers that the key parameter describing a unit of resource belongs to the time domain, namely a time slot for the eMBB services and a mini-slot for the URLLC services. In [19], Anand *et al.* used a model-based approach and proposed linear, convex, and threshold models to capture the eMBB rate loss. Then, they used a gradient-based algorithm to optimize the puncturing placement policies. In [20], Yin *et al.* were interested in the problem of fairness among eMBB users and they proposed a sequential scheduling mechanism which begins by allocating resources to eMBB flows on a slot basis then passes this allocation to the URLLC scheduler which chooses the resources on a mini-slot basis. In [21], Ning *et al.* dealt instead with the problem of reliability of URLLC services when allocating resources, in particular with regard to the allocation of frequency channels and the choice of resources to be punctured. In this same context of reliability, Ben Khalifa *et al.* proposed a channel allocation algorithm for URLLC flows which allows parallel transmissions for the same packet [22]. Similarly, Librino and Santi used a graph-theoretical approach to allocate frequency channels to URLLC flows under interference constraints [23].

From a completely different perspective, Huang *et al.* used a deep reinforcement learning (DRL)-based solution to minimize the negative effect of puncturing on eMBB services [24]. They proposed a series of approximations to circumvent the excessive convergence slowness of the DRL learning phase. Finally, a mixed approach was proposed by Alsenwi *et al.* in [25]. First, they used a model-based decomposition technique to make the resource allocation decision for the eMBB traffic. Then, they used a DRL-based algorithm to optimize the puncturing for the URLLC traffic. The major problem of all these solutions is the attribution of an almost strict priority to the URLLC services, which could harm eMBB flows having rather demanding QoS constraints, or even cause starvation problems for these flows when the URLLC traffic demand is quite high.

The second category aims to make the best use of the flexible rectangular structure of the NR frame to optimize the 2D resource allocation. Model-based and machine learning (ML)-based approaches are also used in this category. In [11], You *et al.* used a model-based approach by proposing an integer linear programming formulation of the packing problem. The objective has been to maximize the throughput of one class of service, e.g. eMBB, in the presence of another class sensitive to delay, e.g. URLLC, under a non-overlapping block constraint. In addition to the over-simplification of the mathematical formulation, the use of a solution approach based on a subgradient method should slow down the resolution. In [26], Sui *et al.* used also a model-based approach and considered the 2D resource allocations that maximize the energy efficiency in the context of flexible numerologies. They defined a quasi-static packing of three types of rectangular resource blocks and used a sliding window based algorithm to find the best allocation for eMBB and URLLC services. The major drawback of this proposal lies in the adoption of this rather rigid approach to block packing. On the other side, Zhang *et al.* took an ML-based approach for TTI selection [12]. First, they used a supervised learning procedure to select a TTI width for each service type. Then, they used a simple binary decision procedure to allocate the resources to each service. In case of non-availability of resources, URLLC flows use a puncturing procedure, while the eMBB flows are queued. Thus, no optimization is applied for the packing of resource

blocks.

B. Fundamental choices and contributions

In the light of the presentation of previous works above, two degrees of freedom are retained in in the design of a resource allocation strategy for heterogeneous services. First, the overall approach: model-based, ML-based, or mixed. Second, the number of dimensions used for the optimization, 1D or 2D, and the scheduling strategy for the heterogeneous services.

It is considered that ML-based approaches currently suffer from three major drawbacks which do not allow them to build a complete, real-time decision-making procedure for flows of different QoS requirements. The first is their too slow convergence which still requires several improvements for the techniques currently in use in both the supervised and unsupervised context. The second is the lack of support for specific QoS constraints for particular applications [25]. In this same context, Bengio *et al.* have recently underlined the inability of these techniques to ensure the feasibility of the solutions obtained and the difficulty of measuring the optimality gap [27]. The last is the use of databases generated by simulations, which neither allows to reliably describe the systems in place, nor to learn the different behaviors of the transmission layers. Notwithstanding this, ML is believed to offer promising solutions for certain specific problems that are worth investigating in the design of future systems. Consequently, the model-based approach is chosen which guarantees a mathematically rigorous description of real phenomena even if certain assumptions are inevitable. Throughout this paper, the parts of our solution will be highlighted where ML can contribute to improving the quality of the solution. This is not to claim that the suggested approach will be hybrid, but it does offer a framework for a hybrid implementation in the future.

For the choice of dimensions, the optimization of resources in 2D is picked which, to the best of the knowledge of the author, has not been considered in previous work and should solve the problem of underutilization of rectangular resources. In addition, the dynamic scheduling approach is to be used which has better resource efficiency than the puncturing approach.

The main contributions in this paper could be summarized as follows:

- Proposing a realistic and standard-compliant architecture for transporting heterogeneous flows from several sources and applications, including sensors and actuators in IoT applications, with different QoS requirements according to the 5G QoS model. Then, the problem is described of jointly mapping the differentiated QoS flows to transport blocks, and packing these blocks into the rectangular time-frequency resource grid.
- Providing a carefully structured mathematical formulation of this problem with the objective of minimizing the overall capacity of the scheduled but not packed transport blocks, and under QoS, capacity and geometrical constraints. Then, two techniques are proposed of decomposition of this joint problem. The first is based on Lagrangian relaxation and requires an iterative solving procedure, while the second is rather intuitive and requires a sequential solving procedure.
- Proposing a configuration-based resolution method that handles most of the processing offline. The proposed approach uses the same fundamental principles of QoS architectures, namely the differentiation of flows and the classification of services. Thus, it should fit easily within the 5G QoS model initially adopted.
- Supported by intensive numerical computations, recommendations are given for an efficient classification of blocks of resources and the quality of the solution provided by the suggested algorithm is compared to the exact solution.

C. Paper organization

This paper is organized as follows: in section II, the basic format of the 5G new radio frame is described and a comprehensive description of the QoS architecture is presented. Then, the resource allocation and packing problems are introduced. In section III, a mathematical formulation is provided of the joint problem of assigning multiple connections to shared transport blocks and then packing these blocks into the rectangular resource grid. In section IV, the suggested configuration-based assignment and packing (CBAP) algorithm is described. An insightful numerical study of the CBAP will be provided in section V. Finally, some concluding remarks are drawn in section VI.

II. 2D RESOURCE ALLOCATION IN 5G NR

In this section, Three things are attempted. First, a brief description is provided of the 2D flexible structure of the 5G new radio (NR) frame [28]. Second, the QoS architecture is presented that will support our system model. Finally, the 2D resource allocation problem is discussed which consists in two steps, the scheduling of flows and the packing of the rectangular resource blocks.

A. Flexible frame structure

The 5G NR frame keeps the same basic structure as the LTE frame with a duration of 10 ms and a breakdown into 10 subframes of 1 ms each. The difference lies in the internal flexibility, in particular with the variation in the capacity of each subframe in terms of *slots*. A slot in 4G has 7 OFDM symbols and lasts 0.5 ms. In NR, a slot has 14 OFDM symbols but its duration varies according to the *numerology* which is defined as the spacing of the subcarriers Δf and the length of the cyclic prefix (CP). Unlike the 4G LTE standard where only one spacing $\Delta f = 15$ kHz is allowed, 5G NR defines a flexible numerology where the spacing of the subcarriers is defined by $\Delta f = 2^{\mu} \times 15$ kHz with $\mu = 0, 1, 2, 3, 4$ [29]. Thus, since the slot always contains 14 symbols, the greater the spectral spacing is, the shorter the duration of the slot becomes, which makes it possible to serve time-sensitive applications. Note also that the NR standard allows the scheduling of *mini-slots* of 2, 4 or 7 symbols, thus further reducing delays and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3074111, IEEE Internet of Things Journal

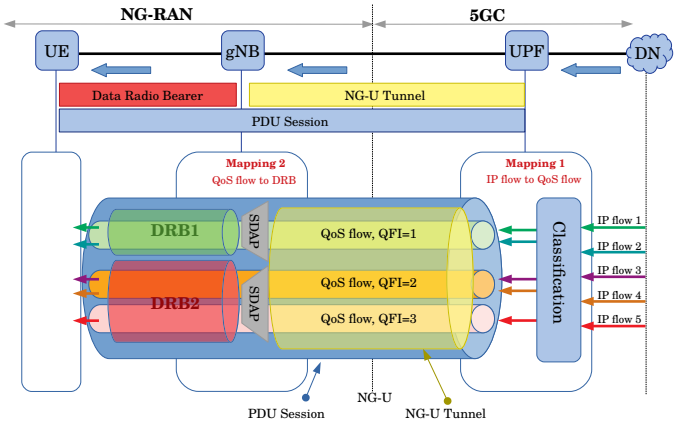

Fig. 1. QoS architecture in 5G NR

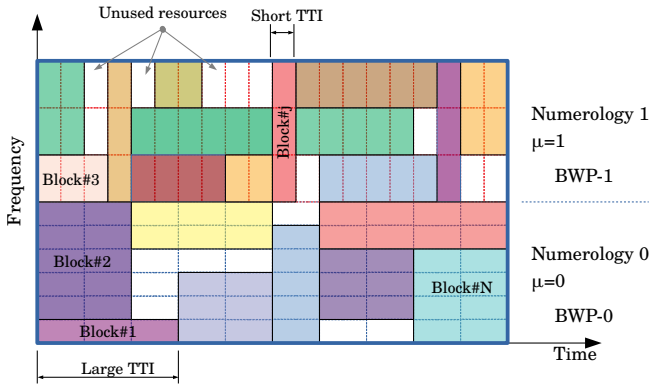

Fig. 2. Flexibility of the NR frame in the time and frequency domains with two numerologies

making it possible to serve applications that are ultra sensitive to delay, such as URLLC [17].

A *physical resource block* (PRB) in NR is composed of 12 subcarriers in the frequency domain, having the same spectral spacing and the same length of CP, over the duration of an OFDM symbol in the time domain. The PRB is the basic allocation unit for physical resources. When several numerologies are supported by the frame, the PRBs of the same numerology form a *resource grid*. Finally, it should be noted that due to the high spectrum width on which 5G operates, the concept of bandwidth part (BWP) was introduced to allow the terminals to operate on specific frequency bands, which saves them energy [17].

B. QoS architecture

The 5G QoS model is based on the concept of *QoS flow* which represents the finest granularity to differentiate the levels of QoS in 5G systems. In fact, all flows associated with the same QoS flow receive the same processing during the transfer. For maximum compatibility with existing standards, the QoS architecture that is to be described here and then used to describe our system model is fully compliant with the 3GPP specifications, e.g. [16] [30]. Fig. 1 summarizes the main concepts associated with the QoS architecture in 5G.

In the downlink direction, when IP data flows from the data network (DN) arrive at the UPF level, they are classified according to packet filter sets and packet detection rules. The corresponding protocol data units (PDU) are marked with a *QoS Flow Identifier* (QFI) which uniquely identifies a QoS flow within a PDU session. Thus, a first mapping, not necessarily one-to-one, is created between the data flows and the QoS flows. Then, the NG-RAN performs a second mapping at the *Service data adaptation protocol* (SDAP) layer between the QoS flows and the DRBs according to the QFI marking and the associated QoS profiles. A DRB transfers packets with the same processing. Separate DRBs can be established for QoS flows requiring different transfer processing.

C. Resource allocation

At the MAC sublayer, the QoS flows encapsulated in DRBs, then in Radio link control (RLC) channels and finally in logical channels must be placed in transport blocks (TB). A transport block represents the MAC PDU which is the concatenation of several data units, each having a specific header. The MAC scheduler must select and multiplex the QoS flows that could be served, and then assign radio resources in a unit of slot, e.g. one mini-slot, one slot or multiple slots, to every scheduled TB.

The description of physical resources presented in the previous section shows that blocks of rectangular resources of different sizes must be allocated to individual or aggregated flows, allowing them to benefit from dynamic TTIs. This rectangular *packing* generally allocates more resources to each user than necessary. This difference is to be called an *overallocation* in the rest of the present paper. In addition, when packing these blocks in the 2D time-frequency frame, it is very likely to leave some empty spaces, called *unused resources*, which are too small to place additional blocks in.

Fig. 2 illustrates some basic principles of the flexible NR frame structure in the time and frequency domains. The use of flexible TTIs makes it possible to adapt the heterogeneous flows in the same frame. The generic term *block* is used to denote a rectangular allocation on the frame. This block could be associated with a user, a group of users, a slice in network slicing architectures [31] or any particular service, e.g., eMBB, mMTC, URLLC, etc.

The problem of placing rectangular blocks in the frame can be classified as a 2D rectangular single large object packing problem which is an intermediate type of the knapsack problem and could also be considered a special case of the 2D bin packing (2BP) [32]. In fact, according to a regular 2BP, a given set of rectangular objects (blocks), with possibly different but well-known widths and heights, must be packed into a minimum number of identical rectangular *bins* (frames) [33]. Thus, the problem of resource allocation in NR can be seen as a particular case of 2BP as only one frame is considered.

III. The Joint Mapping and Packing Problem

A. Problem description and formulation

A single cell wireless network is considered where a next-generation base stations called gNB serves in the downlink

TABLE I
SUMMARY OF VARIABLES AND PARAMETERS

INDEX	*Description*
k	Indexing for user equipment UE
(i,k)	Indexing for QoS flow
p	Indexing for TB class
q	Indexing for QoS flow type
SETS	*Description*
K	Total number of UEs in the system
N_k	Total number of QoS flows for user UE$_k$
$\mathcal{P}$	Total number of block classes
$\mathcal{Q}$	Total number of QoS flow types
M^p	Max number of TBs of class p that can be packed
VARIABLES	*Description*
$\gamma_{i,j}^{p,k}$	Mapping decision of QoS flow (i,k) to TB (j,p,k)
$\delta_j^{p,k}$	Scheduling decision of TB (j,p,k)
$\pi_j^{p,k}$	Packing decision of TB (j,p,k)
l_{js}	TB (j,p,k) is placed on the left of TB (s,p',k')
u_{js}	TB (j,p,k) is placed below TB (s,p',k')
$w_j^{p,k}$	Width of rectangular block (j,p,k)
$h_j^{p,k}$	Height of rectangular block (j,p,k)
$(x_j^{p,k}, y_j^{p,k})$	Lower left coordinates of rectangular block (j,p,k)
$\alpha_{i,j}^{p,k}$, $\lambda_j^{p,k}$, β	Lagrange multipliers
$z_l^{p,k}$	Number of TB configuration l in the class p for UE$_k$
PARAMETERS	*Description*
W	Width of the resource grid (symbols)
H	Height of the resource grid (PRBs)
b_i^k	Bandwidth request of QoS flow (i,k)
b_q	Bandwidth of demands of type q
r_i^k	Minimum robustness of QoS flow (i,k)
r_q	Minimum robustness of demands of type q
C^p	Transport block size of class p
R^p	Spectral efficiency of class p
$\bar{R}$	Best modulation efficiency that can be achieved
$\boldsymbol{e}_l^p$	The l^{th} configuration of a TB of class p
$e_{q,l}^p$	Number of QoS flows of type q in the configuration $\boldsymbol{e}_l^p$
ψ_l^p	Cost of a TB of configuration l in class p

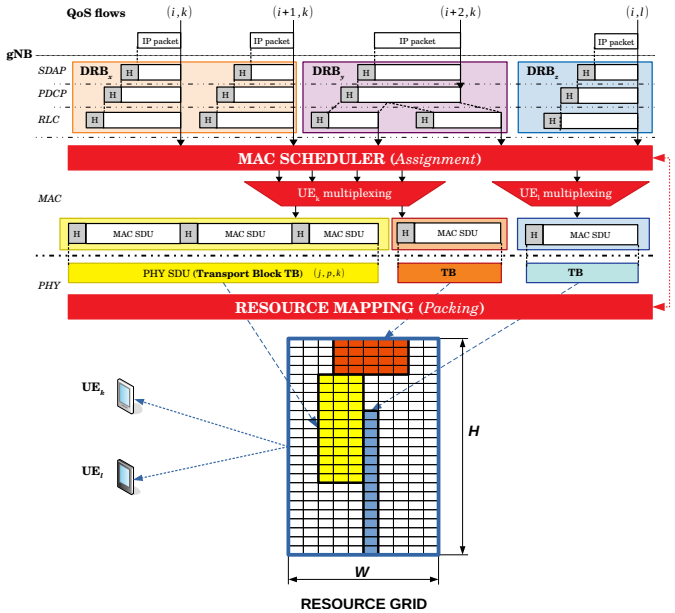

Fig. 3. Problem description

direction K user equipments (UE) of several types requiring different levels of QoS depending on the nature of the services they use, e.g. eMBB, URLL, mMTC, IoT, etc. It is assumed that a classifier at the edge of the 5GC has already identified the different IP flows according to well-defined classification rules and marked each packet with a QFI identifier as shown in Fig. 1. When the set of all QoS flows thereby formed arrive at the data link layer of the gNB, a scheduling decision must be made to determine the list of flows likely to be served in the next 10 ms NR frame. However, since the scheduling decision is made for each subframe, the joint problem over the duration of a single subframe of 1 ms will be formulated and solved. In addition, the resource grid described in Fig. 3 will be considered as the basic unit of an NR subframe and thus the resolution will be limited to the borders of this area. A summary of variables and parameters is provided in Table I.

Each UE$_k$, $k = 1,\ldots,K$ must receive N_k QoS flows that are identified with the pair (i,k), $i = 1,\ldots,N_k$. Each QoS flow (i,k) is characterized by a bandwidth request b_i^k expressed in number of bits per ms and a minimum level of robustness r_i^k expressed in bits per resource element (RE). The robustness of the modulation and coding scheme (MCS) is determined from the channel quality indicator (CQI) sent by the UE.

These QoS flows must be mapped to TBs. Obviously, it is not possible to know in advance the TBs necessary for each UE, which does not make possible the resolution of this mapping problem with the usual assignment algorithms. However, it could be assumed that TBs are listed in several classes which can be associated with services like video streaming, voice over IP and data, with service categories like eMBB, URLLC or mMTC, or any other classification criteria. Each *class* $p = 1,\ldots,\mathcal{P}$, is defined by a transport block size C^p expressed in units of PRBs and an MCS which provides a spectral efficiency R^p in bits per RE. The j^{th} TB of class p assigned to the UE$_k$ is identified by the triplet (j,p,k). Finally, given the capacity of the grid in PRBs, it is possible to determine the maximum number M^p of TBs of class p that can be placed.

A second step consists in packing the rectangular resource blocks of variable sizes associated with the selected TBs in the resource grid. Fig. 3 shows an example of the construction of transport blocks from IP packets belonging to different QoS flows and intended for 2 UEs k and l. The first TB noted (j,p,k) groups together data units coming from DRBs x and y. The second TB contains a packet fragment from DRB y. Finally, the third TB is constructed from data intended for UE$_l$. Thus, it can be seen that several criteria can intervene in the construction of TBs. Obviously, some blocks may not be packed.

In order to formulate this problem, three series of binary decision variables are used. In the first set, the mapping variable $\gamma_{i,j}^{p,k}$ is used to indicate whether QoS flow (i,k) is mapped to TB (j,p,k) and the TB scheduling variable $\delta_j^{p,k}$ to indicate whether TB (j,p,k) is scheduled. In the second set, the packing variable $\pi_j^{p,k}$ is used to indicate whether the block associated with TB (j,p,k) is packed in the grid. Finally,

the third set consists of the 2D packing variables. Since the linear modeling technique described in [34] is going to be used here, the l_{js} variable is to be used to indicate whether the block associated with TB (j,p,k) is placed on the left of TB (s,p',k') and the variable u_{js} to indicate whether the block associated with TB (j,p,k) is placed below TB (s,p',k').

Furthermore, it is assumed that each rectangular block (j,p,k) associated with TB (j,p,k) is geometrically defined by its width $w_j^{p,k}$ expressed in number of symbols, its height $h_j^{p,k}$ expressed in number of PRBs and its lower left coordinates $(x_j^{p,k}, y_j^{p,k})$ which are also variables of the packing problem. To make the notation less crowded, the indices p of the class and k of the UE are removed in the relative position variables l_{js} and u_{js}.

The resource grid is characterized by a width $W = 14 \times 2^{\mu}$ expressed in number of symbols and a height $H = N_{RB}^{\mu}$ expressed in number of PRBs.

The objective of the overall problem is to minimize the total capacity of TBs scheduled but not packed due to lack of radio resources. The problem is formulated as follows:

$$\min \sum_{k=1}^{K} \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} C^p(\delta_j^{p,k} - \pi_j^{p,k}) \tag{1}$$

subject to

$$\sum_{j,p} \gamma_{i,j}^{p,k} = 1 \qquad \forall (i,k) \tag{2}$$

$$\sum_{i=1}^{N_k} \left\lceil \frac{b_i^k}{12.R^p} \right\rceil \gamma_{i,j}^{p,k} \leq C^p \qquad \forall (j,p,k) \tag{3}$$

$$r_i^k \gamma_{i,j}^{p,k} \geq R^p - (1 - \gamma_{i,j}^{p,k})\bar{R} \qquad \forall (i,k) \tag{4}$$

$$\delta_j^{p,k} - \gamma_{i,j}^{p,k} \geq 0 \qquad \forall (i,k)\ \forall (j,p,k) \tag{5}$$

$$\delta_j^{p,k} - \pi_j^{p,k} \geq 0 \qquad \forall (j,p,k) \tag{6}$$

$$w_j^{p,k} h_j^{p,k} \geq C^p \qquad \forall (j,p,k) \tag{7}$$

$$\sum_{(j,p,k)} C^p \pi_j^{p,k} \leq HW \tag{8}$$

$$l_{js} + l_{sj} + u_{js} + u_{sj} + (1 - \pi_j^{p,k}) + (1 - \pi_s^{p',k'}) \geq 1 \qquad \forall (j,p,k),(s,p',k') \tag{9}$$

$$(x_j^{p,k} + w_j^{p,k}) - x_s^{p',k'} \leq W(1 - l_{js}) \quad \forall (j,p,k),(s,p',k') \tag{10}$$

$$(y_j^{p,k} + h_j^{p,k}) - y_s^{p',k'} \leq H(1 - u_{js}) \quad \forall (j,p,k),(s,p',k') \tag{11}$$

$$x_j^{p,k} + w_j^{p,k} \leq W + (1 - \pi_j^{p,k})W \qquad \forall (j,p,k) \tag{12}$$

$$y_j^{p,k} + h_j^{p,k} \leq H + (1 - \pi_j^{p,k})H \qquad \forall (j,p,k) \tag{13}$$

$$l_{js},\ u_{js} \in \{0,1\} \qquad \forall (j,p,k),(s,p',k') \tag{14}$$

$$w_j^{p,k},\ h_j^{p,k},\ x_j^{p,k},\ y_j^{p,k} \geq 0 \quad \text{and integer} \quad \forall (j,p,k) \tag{15}$$

$$\gamma_{i,j}^{p,k},\ \delta_j^{p,k},\ \pi_j^{p,k} \in \{0,1\} \qquad \forall (i,k)\ \forall (j,p,k). \tag{16}$$

Constraints (2) ensure that each flow is assigned to exactly one TB.

Constraints (3) limit the total flows assigned to a TB of class p to the capacity of this class. A simple and intuitive formula was used to convert the target bit rate b_i^k to PRBs as described in [35]. In fact, since a PRB contains 12 REs and each RE carries a number of bits according to the chosen modulation order defined by R^p, the resource requirements of a QoS flow can be directly obtained if it is transported on a TB of class p.

It is worth noting that a more insightful way to express this capacity constraint would be to consider the achievable rate instead of the target rate. The achievable rate of user UE_k on TB (j,p,k) could be defined as a function $\varphi(h_j^{p,k}, y_j^{p,k}, \bar{\Gamma}_j^{p,k})$, where $\bar{\Gamma}_j^{p,k}$ is the average signal-to-interference plus noise ratio (SINR) [36]. This would add two more levels of complexity to our problem. First, the additional dependence between the scheduling problem and the packing problem can be noticed through the variables h and y, which indicate the bandwidth of the TB and the index of the first PRB on the frequency axis, respectively. Second, the average SINR $\bar{\Gamma}_j^{p,k}$ should introduce a new variable indicating the amount of power allocated to QoS flow (i,k) if it is transported on TB (j,p,k). Consequently, our model should include an additional margin adaptive optimization which consists in minimizing the overall power consumption under the following constraint [37], [38]:

$$\varphi(h_j^{p,k}, y_j^{p,k}, \bar{\Gamma}_j^{p,k}) \geq b_i^k \tag{17}$$

Ultimately, four problems will have to be jointly solved: our two initially defined scheduling and packing problems, plus two other new problems of subchannel assignment and power allocation, which would be too complex. Thus, only formulation (3) will be considered in the rest of this paper.

Constraints (4) ensure that each connection receives at least the MCS which provides the minimum required efficiency. A flow can be assigned to a TB offering a more robust modulation than it requires. This kind of assignment could be beneficial if it contributes to the reduction of control traffic or better packing. Obviously, a flow cannot be assigned to an unused TB as indicated in constraints (5). Note that the variables $\delta_j^{p,k}$ are intermediate variables which relate the assignment variables $\gamma_{i,j}^{p,k}$ to the packing variables $\pi_j^{p,k}$ through constraints (5) and (6). Constraints (6) indicate that it is not always possible to place all the TBs used in the first step in the time-frequency resource grid. Constraints (7) allow the freedom to choose the dimensions of the physical resource blocks associated with each TB. Constraint (8) fixes the limit of total capacity of the grid in number of PRBs.

For 2D packing, constraints (9, 10 and 11) ensure that no overlap occurs between the blocks, and constraints (12 and 13) prevent the crossing of the grid borders.

B. Problem decomposition

In addition to the high computational complexity of this non-linear mixed integer programming problem (MINLP), it should be noted that the interdependence between the solution of the assignment problem and the solution of the packing problem makes the resolution of the overall problem with the usual MINLP solvers quite difficult. However, decomposition techniques can be used to make the problem more tractable.

Our first decomposition uses the Lagrangian relaxation technique [39]. By examining the different constraints of this

problem, three coupling constraints are identified. Constraints (5) and (6) connect the different decision variables and constraint (8) links the decisions made for all users. The present study will, therefore, proceed to relaxing these constraints by adding three sets of non-negative Lagrange multipliers: $\alpha_{i,j}^{p,k}$ for constraints (5); $\lambda_j^{p,k}$ for constraints (6); and β for constraint (8).

The Lagrangian could be expressed as follows:

$$\mathcal{L} = \min \sum_{k=1}^{K} \left\{ \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} \left[\left(C^p - \lambda_j^{p,k} - \sum_{i=1}^{N_k} \alpha_{i,j}^{p,k} \right) \delta_j^{p,k} - \left(C^p - \lambda_j^{p,k} - \beta C^p \right) \pi_j^{p,k} + \sum_{i=1}^{N_k} \alpha_{i,j}^{p,k} \gamma_{i,j}^{p,k} \right] \right\} - \beta HW \tag{18}$$

$$\text{subject to constraints} \quad (2, 3, 4, 7) \text{ and } (9 - 16).$$

The first consequence of relaxation is obtaining a separate problem for each UE, which makes it possible to solve subproblems of dimensions much smaller than that of the original problem. The second consequence is the decomposition of each subproblem for a UE_k into 3 independent subproblems. The first is a subproblem for assigning QoS flows to TBs, which will be called F2BAP$^{(k)}$. This subproblem is expressed by the following formulation.

$$\min_{\gamma} \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} \sum_{i=1}^{N_k} \alpha_{i,j}^{p,k} \gamma_{i,j}^{p,k} \tag{19}$$

$$\text{subject to constraints} \quad (2, 3, 4).$$

The second is a 2D packing subproblem, which will be called 2DBPP$^{(k)}$. This subproblem is formulated as follows:

$$\min_{\pi} \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} \left((\beta - 1) C^p + \lambda_j^{p,k} \right) \pi_j^{p,k} \tag{20}$$

$$\text{subject to constraints} \quad (9 - 15).$$

Finally, the third subproblem, which will be called SP3$^{(k)}$, consists of an unconstrained minimization of the following function:

$$\min_{\delta} \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} \left(C^p - \lambda_j^{p,k} - \sum_{i=1}^{N_k} \alpha_{i,j}^{p,k} \right) \delta_j^{p,k} \tag{21}$$

whose solution is straightforward. If $\left(C^p - \lambda_j^{p,k} - \sum_{i=1}^{N_k} \alpha_{i,j}^{p,k} \right) \leq 0$, then $\delta_j^{p,k} = 1$; otherwise, $\delta_j^{p,k} = 0$.

After identifying the various subproblems, a subgradient method can be applied to find an exact solution to the original problem. This method can be summarized in the following iterative procedure:

1: Initialize the Lagrange multipliers: $\alpha_{i,j}^{p,k}$, $\lambda_j^{p,k}$, and β.
2: **repeat**
3: For all k, solve subproblems F2BAP$^{(k)}$, 2DBPP$^{(k)}$, and SP3$^{(k)}$ ▷ in parallel
4: Update the Lagrange multipliers
5: **until** stopping criterion is reached

The major difficulty of this iterative solution approach lies in the initialization and updating of Lagrange multipliers. The relatively slow convergence of this method makes it unattractive for decision making in real-time.

Our second decomposition is rather intuitive and is based on an inspection of the structure of the problem. Indeed, apart from the coupling constraints of the decision variables (6), the problem can be broken down into two separate subproblems, one for the scheduling decision and the other for block packing.

The modified QoS flow to TB mapping subproblem F2BAP' is formulated as follows:

$$\min_{\gamma,\delta} \sum_{k=1}^{K} \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} \sum_{i=1}^{N_k} C^p \delta_j^{p,k} \tag{22}$$

$$\text{subject to constraints} \quad (2 - 5)$$
$$\gamma_{i,j}^{p,k},\ \delta_j^{p,k} \in \{0, 1\}.$$

Similarly, the modified subproblem 2DBPP' is formulated as follows:

$$\max_{\pi} \sum_{k=1}^{K} \sum_{p=1}^{\mathcal{P}} \sum_{j=1}^{M^p} C^p \pi_j^{p,k} \tag{23}$$

$$\text{subject to constraints} \quad (7 - 15)$$
$$\pi_j^{p,k} \in \{0, 1\}.$$

The question which arises now is to know how one can bypass constraint (6) to obtain a solution to the original problem. Actually, it suffices to solve these two subproblems sequentially starting with the mapping subproblem F2BAP', then knowing the blocks that have been built, solve the packing subproblem 2DBPP'. The solution obtained is necessarily approximate but it will be shown in the rest of this paper that it is very close to the optimal solution. Moreover, an efficient algorithm will be proposed in the following section that should make it possible to solve these subproblems within an acceptable time frame.

IV. Configuration-based joint assignment and packing

In the previous section, several *classes* of TBs were defined. Likewise, it will be assumed here that QoS flows can also belong to specific types. Each *type*, $q = 1, \ldots, \mathcal{Q}$, defines a specific application, for example, traffic from IoT, VoIP or video streaming, which requires fixed bandwidth b_q and a minimum level of robustness r_q. In the NR architecture, this classification is carried out by the SDAP layer and gives rise to data radio bearers. Thus, it will be assumed in what follows that QoS flows are assigned to DRBs whose types are predefined.

A. Configuration-based assignment

The assignment problem could be redefined as follows. The set $\mathcal{F}$ of QoS flows will be considered again and it will be assumed that each QoS flow can be described with a type q.

TABLE II
EXAMPLE OF THE TYPES OF DATA FLOWS

q	Application	b_q (bytes/ms)	r_q (bytes/PRB)
1	VoIP (G.711)	8	3 *(QPSK)*
2	VoIP (G.711)	8	9 *(QAM64)*
3	Video (H.264)	60	3 *(QPSK)*
4	Video (H.264)	60	9*(QAM64)*

Therefore, the data flow requests can be presented as a request vector $\boldsymbol{n}^k = (n_1^k, \ldots, n_q^k, \ldots, n_{\mathcal{Q}}^k)$ where n_q^k is the number of QoS flows of type q intended for UE_k. It will also be assumed that a set $\mathcal{P}$ of classes of TBs is given. The minimum number of TBs of each class to be used needs to be determined in order to assign all data flows.

Taking advantage of the fact that the QoS flow types and TB classes are completely known, the present work will opt for an *offline* solution technique for this assignment problem. For each class p of TB, an exhaustive list will be made of all possible assignments of the types of QoS flows. Each assignment is called a *configuration* of a TB. The l^{th} configuration of a TB of class p is defined as a vector $\boldsymbol{e}_l^p = (e_{1,l}^p, \ldots, e_{q,l}^p, \ldots, e_{\mathcal{Q},l}^p)$ where $e_{q,l}^p$ is the number of QoS flows of type q which could be assigned to a TB of class p.

For example, it is supposed that a class p is defined with a capacity $C^p = 20$ PRBs and an efficiency $R^p = \frac{12\times 4}{8} = 6$ bytes/PRB corresponding to a QAM16 modulation which allows 4 bits to be transmitted by one RE. It is also assumed that the requests belong to four distinct types as summarized in Table II. In order to list all the possible configurations for this TB class, the requested bandwidths must be first converted into PRBs using the modulation efficiency of the class. In this example, VoIP streams consume $\left\lceil \frac{8}{6} \right\rceil = 2$ PRBs, and video streams $\left\lceil \frac{60}{6} \right\rceil = 10$ PRBs. Second, the types that cannot be served by the TB are to be filtered because they require a higher level of robustness. In this example, types 1 and 3 cannot be mapped to this class due to the robustness constraint. Finally, the following configurations are obtained:

$$
\begin{aligned}
\boldsymbol{e}_1^p &= (0,1,0,0)\\
\boldsymbol{e}_2^p &= (0,2,0,0)\\
&\ldots\\
\boldsymbol{e}_{10}^p &= (0,10,0,0)\\
\boldsymbol{e}_{11}^p &= (0,0,0,1)\\
\boldsymbol{e}_{12}^p &= (0,1,0,1)\\
\boldsymbol{e}_{13}^p &= (0,2,0,1)\\
&\ldots\\
\boldsymbol{e}_{16}^p &= (0,5,0,1)\\
\boldsymbol{e}_{17}^p &= (0,0,0,2).
\end{aligned}
$$

Each configuration l of a TB of class p satisfies the TB capacity constraint and the transmission robustness constraints (24 and 25).

$$\sum_{q=1}^{\mathcal{Q}} e_{q,l}^p \left\lceil \frac{b_q}{R^p} \right\rceil \leq C^p \tag{24}$$

$$(r_q - R^p) e_{q,l}^p \geq 0 \qquad \forall\, q \in \mathcal{Q}. \tag{25}$$

It is also to be noted by constraint (25) that a data stream can have a more robust transmission than that requested. In fact, it is not necessary to define a TB class for all possible levels of transmission robustness. In addition, even if a TB class is defined for a required level of robustness, it might sometimes be more advantageous to map the data flow corresponding to a TB which provides more robust transmission if this could minimize the number of TBs. This allocation in a more robust class will be called a *service migration*.

After defining the relevant TB classes and data flow types, it is possible to create an *offline* database called TBDB of all TB configurations. Therefore, the only *online* processing to do for each request vector $\boldsymbol{n} = \left[\boldsymbol{n}^1, \ldots, \boldsymbol{n}^K\right]$ is to find the optimal number of configurations that support all QoS flows. Let $z_l^{p,k}$ be an integer variable which determines the number of TB configuration l in the class p for UE_k. The Configuration-Based Assignment (CBA) problem could be formulated as a linear integer problem:

$$\min \sum_k \sum_p \sum_l \psi_l^p z_l^{p,k} \tag{26}$$

$$\sum_p \sum_l e_{q,l}^p z_l^{p,k} = n_q^k \tag{27}$$

$$z_l^{p,k} \geq 0 \quad \text{and integer}, \tag{28}$$

where ψ_l^p is the cost of a TB of configuration l in class p. It is important to note that the CBA problem is also separable by UE.

A possible cost of a TB is the additional overhead it can cause in the grid. This cost will depend on the number of connections $e_{q,l}^p$ of each type q in the configuration. However, if a cost factor is chosen that does not depend on the number of connections, e.g. the capacity of the class in number of PRBs, the costs of all the configurations of the same class will be equal. This implies that in the above example, the configurations $\boldsymbol{e}_1^p$ to $\boldsymbol{e}_{16}^p$ will have the same cost. More specifically, if only VoIP connections are considered, i.e. $\boldsymbol{e}_1^p$ to $\boldsymbol{e}_{10}^p$, putting 1 or 10 connections will result in the same cost. Similarly, for $\boldsymbol{e}_{11}^p$ to $\boldsymbol{e}_{16}^p$, if a video connection is mapped, 0 to 5 VoIP connections are possible to add without incurring any additional costs. Consequently, when the cost depends only on the class of TB, it is not necessary to make a complete enumeration of all the possible configurations. In fact, it suffices to enumerate only the configurations that fully fill the capacity of the TB; called *maximum configurations* here, $\bar{\boldsymbol{e}}_l^{p,k}$. The size of the TBDB database of TB configurations will be considerably reduced, as will the execution of the linear model (26 – 28). The equality constraint (27) must be transformed into inequality as follows:

$$\sum_p \sum_l \bar{e}_{q,l}^p z_l^{p,k} \geq n_q^k. \tag{29}$$

this second approach will be chosen with maximum configurations and will be discussed in section V three possible configuration cost functions.

Finally, note that a mixed-integer linear programming (MILP) solver, i.e. CPLEX Optimizer, will be used to solve the CBA problem. However, it would be beneficial to investigate in future work ML techniques for solving such a combinatorial optimization problem [27].

B. Configuration-based packing

Taking advantage of the prior knowledge of the capacities C^p of the TB classes, it is possible to solve most of the packing problem offline as well. Indeed, an enumeration can be made of all the possible configurations of each grid in a specific numerology and BWP, as well as create an offline database of *grid configurations* to be called (GDB). The i^{th} configuration of a grid is defined as a vector $\boldsymbol{\tau}_i = (t_i^1, \ldots, t_i^p, \ldots, t_i^{\mathcal{P}})$ where t_i^p is the number of TBs of class p which could be packed in the grid.

However, it is relatively difficult to make an exhaustive list of all possible grid configurations since it is not easy to predict the amount of overallocation. Indeed, even if the sum of the capacities of the QoS flows in the request vector $\boldsymbol{n}$ should not exceed the grid capacity, an overallocation can reasonably be inevitable because the CBA problem must affect all the flows to the predefined TB classes, as it is explicitly stated in constraint (29). Therefore, since only the TB capacity in number of PRBs must be considered when listing the grid configurations, just the following constraint must be satisfied

$$\sum_p t_i^p C^p \leq HW + \Omega \tag{30}$$

where Ω is the overallocation. Each grid configuration is used as an entry for the 2DBPP' problem to find an exact solution for its packing. Obviously, some TBs might not be packed because of the geometrical constraints.

A way to set a value of Ω will be provided in section V via numerical studies. Nevertheless, a procedure to design the GDB as an incremental database will be described in the next subsection a procedure. The database is initialized with a complete enumeration of grid configurations for a minimum value of Ω, then the new configurations are added as they occur. ML techniques could also be used to find the subset of relevant grid configurations given a traffic demand pattern.

C. Description of the CBAP Algorithm

The two major steps of our configuration-based assignment and packing (CBAP) algorithm are described in the flowchart of Fig. 4. The first is the offline processing where we build the TB and grid databases. The second is the online processing when a vector request of categorized QoS flows is presented to the MAC scheduler. As we can see, the main online treatment is limited to the resolution of the CBA problem. Let us recall that since the objective of this work is to show the relevance of the whole resolution procedure both at offline and online levels, an exact resolution procedure for the CBA problem using the CPLEX solver will be used. The proposal for approximate

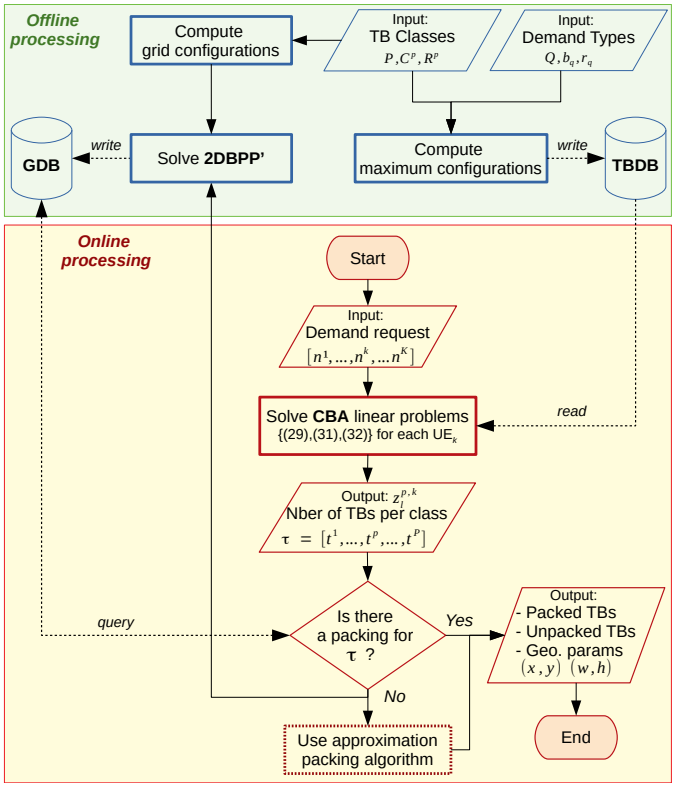

Fig. 4. Flowchart of the CBAP algorithm

procedures potentially based on ML will be the subject of future investigations. From the solution vector $\boldsymbol{z}$ of the CBA, a new vector $\boldsymbol{\tau}$ is built which considers only the number of required blocks in each class using $t_i^p = \sum_k \sum_l z_l^{p,k}$. The packing problem is reduced to a simple database query to determine which block could be packed and where it should be placed. However, if the query does not return any record, a simple approximation packing algorithm could be used, e.g. eOCSA [40] for the current request and submit the vector $\boldsymbol{\tau}$ for an offline processing to get an exact solution with the 2DBPP' and record it for future use in the GDB.

If it is assumed that the GDB contains the requested packing, the complexity of the CBAP algorithm is directly related to the resolution of the CBA integer linear problems for each UE$_k$. Since the $\bar{e}_{q,l}^p$ coefficients are also integers and the size of the TBDB is fixed, then these problems are polynomially solvable [41]. Finally, no extra signaling overhead is expected for the implementation of the CBAP algorithm. Each UE should receive its resource grants on the signaling channel as expected.

V. Numerical Results

In this section, some decision criteria are first provided for choosing the key parameters of our joint approach to flow assignment and block packing described by the CBAP algorithm. Next, the performance of the configuration-based assignment algorithm is compared to the optimal solution of subproblem F2BAP'. The resolution of subproblems F2BAP' and 2DBPP', as well as of the linear problem CBA for the online search of the configurations are made with CPLEX

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3074111, IEEE Internet of Things Journal

12.6.0 on a 64-bit Windows 10 PC with a 2.7GHz Intel Core i7-7500U CPU, 8GB of RAM.

A resource grid is considered in a 10 MHz channel and in the numerology $\mu = 0$. Chosen are $H = N_{RB}^{\mu} = 30$ PRBs and obviously $W = 14$ symbols. A set of 420 PRBs available for the exchange of data traffic and control will be obtained. According to TS38.306 [35] which provides a formula for calculating the useful bit rate, the overload on the downlink is evaluated at 14%, the equivalent of roughly 60 PRBs or two columns on the resource grid. It will, therefore, be assumed that the control traffic is grouped together at the start of the subframe and only a grid is considered of useful resources of width $W = 12$ and a total capacity of 360 PRBs.

All results are averaged over 100 demand vectors, and each demand vector contains 100 connections on average. With a total amount of resources around 360 PRBs, it can be said that serving 100 connections is a sufficiently high level of scale.

A. Types and classes

The suggested approach is based on two fundamental choices, namely the QoS flow types and the TB classes. Although our CBAP algorithm allows a free definition of types that can cover several applications depending on the service provider and potentially all the MCSs authorized by the standard, the definition of classes is quite critical for performance. In the following, a discussion will be provided of the key performance criteria and some recommendations for an effective definition of the classes.

1) Types of demands: First, types are defined according to three typical services with heterogeneous requirements: multimedia streaming (voice/video), data traffic, and IoT applications. For the streaming application, a medium resolution of 640x360 has been chosen which gives a typical bitrate between 800 and 1200 kbps for advanced terminals [42]. For the data application, the elastic nature of this traffic is exploited to enforce heterogeneity. Finally, the bitrate of the IoT application rather reflects the size of a transaction [43]. Three MCSs are defined which satisfy three completely different robustness requirements: QPSK providing only 3 bytes/PRB for remote users, QAM16 providing 6 bytes/PRB for users in the middle of the cell and QAM64 providing 9 bytes/PRB for users near the base station. The resulting types are described in Table III and are indexed in ascending order by the number of required PRBs.

2) Classes of transport blocks: In order to fix the set of classes of TBs, four sets of profile vectors are selected: V, W, X and Z. Each set is composed of five capacity vectors of six classes with increasing standard deviation values. This is not to pretend that the standard deviation will be the most decisive parameter, but that it will be chosen to set out certain performance criteria. The set of vectors V, as well as the number of maximum configurations they generate are represented in Table IV. The details and the configurations of all vectors are provided in IEEE DataPort [44].

It is worth noting that the size of the TBDB increases exponentially with the standard deviation when service migration is being used. This same behavior is observed for the other three sets, as shown in Fig. 5 which indicates the average sizes of all the sets. Whether it is worth using service migration and incurring its greater complexity or not is to be discussed below.

TABLE III
TYPES OF REQUESTS USED IN THE SIMULATION

Flow		**Types**		
Application	Bitrate $(bits/ms)$	Index (i)	Minimum requirement (PRB)	Efficiency (bytes/PRB)
Smart grid	72	1	**1**	9
		2	**2**	6
		3	**3**	3
Data	288	4	**4**	9
		5	**6**	6
		6	**12**	3
Streaming	864	7	**12**	9
		8	**18**	6
		9	**36**	3

TABLE IV
PROFILES OF THE TBS

Vector	**Classes:** $p = 1, \ldots, 6$						**Statistics**		**TBDB size**	
	1 QAM64	2 QAM16	3 QPSK	4 QAM64	5 QAM16	6 QPSK	Mean	Std Dev	mig	w/o
V_A	14	22	27	36	39	42	30	10	924	52
V_B	9	14	25	38	42	52	30	15	1507	53
V_C	6	10	18	36	50	60	30	20	2528	55
V_D	4	8	12	36	45	75	30	25	4829	55
V_E	2	5	8	28	52	85	30	30	8506	51

B. Cost of the configurations

In this study, we compare three possible configuration costs ψ_l^p. First, the same linear coefficient $\psi_l^p = C^p$ is used as in the objective-function (1). However, in this case, a configuration of capacity 8 will have the same cost as 4 configurations of capacity 2. Thus, the linearity should increase the number of TB blocks. Second, a logarithmic coefficient $\psi_l^p = log_2 C^p$ is used in order to minimize the number of TBs and to encourage the use of larger blocks. Third, a penalty ϵ is added to the cost $\psi_l^p = C^p + \epsilon$, so that a block of capacity 8 is preferred to 4 blocks of capacity 2. Indeed, 4 blocks of 2 PRBs will cost 8 without penalties, which is the same cost as a block of 8 PRB, while their cost with penalty will be $4 \times (2 + \epsilon) = 8 + 4\epsilon$

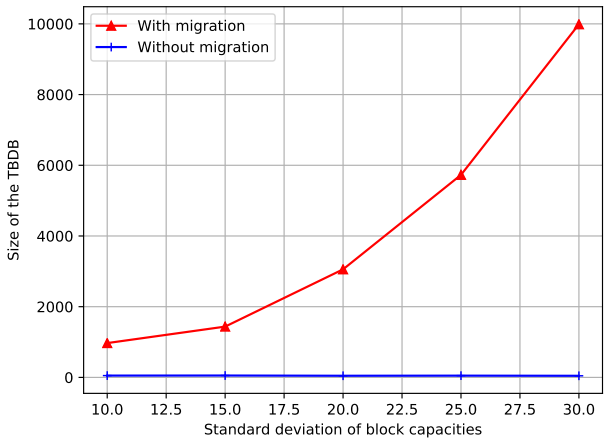

Fig. 5. Size of the database TBDB

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3074111, IEEE Internet of Things Journal

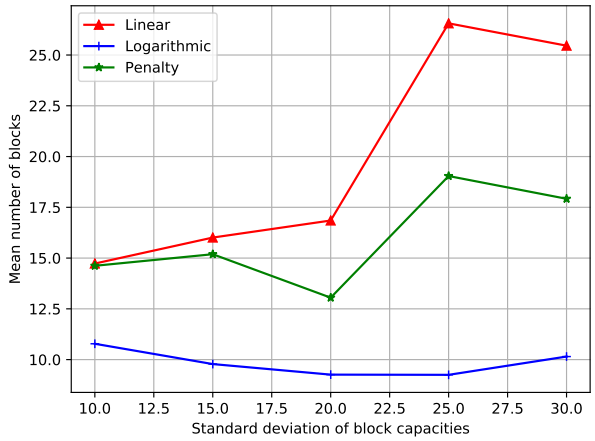

Fig. 6. Number of blocks generated by different configuration costs

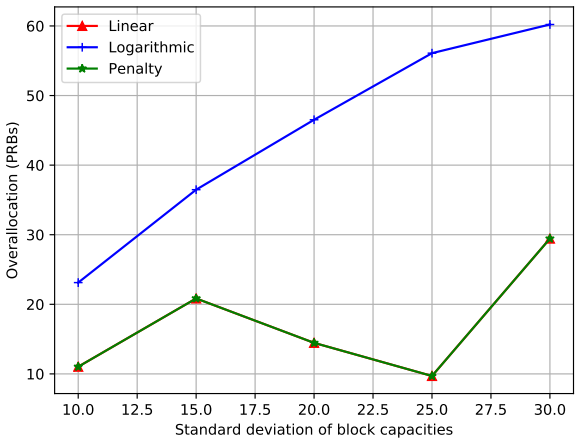

Fig. 7. Overallocation generated by different TB cost functions

which is greater than the cost of one single block of 8, which is only $8 + \epsilon$ in this case. ϵ is taken to be equal to 0.1.

Fig. 6 reports the number of blocks incurred by each cost coefficient. As expected, the linear cost generates the greatest number of blocks, especially for a higher standard deviation. The logarithmic

In Fig. 6, we report the number of blocks incurred by each cost coefficient. As expected, the linear cost generates the greatest number of blocks, especially for a higher standard deviation. The logarithmic coefficient generates a much lower number of blocks. In terms of overallocation, which was defined earlier as the number of PRBs added in order to send the data in rectangular form, it can be noted in Fig. 7 that the linear cost and penalty coefficients provide exactly the same performance, which is clearly better than the logarithmic coefficient. Indeed, the logarithmic coefficient encourages the use of large capacity blocks that may not be completely filled. It can be concluded that the cost with penalty provides an acceptable compromise between the size of the TBCB and the performance. Therefore, according to this first survey, only the penalty-based cost will be used in the rest of this section.

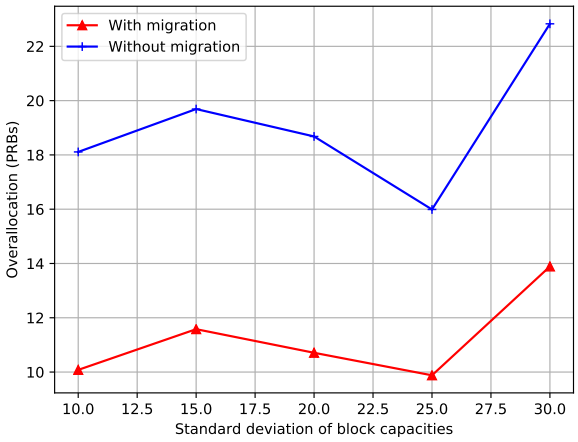

Fig. 8. Performance of CBA in terms of overallocation

C. Standard-deviation of the class capacity

In this study, the best standard deviation is sought for the class capacity values of TBs. First, the performance of the CBA is assessed in terms of overallocation and the number of TBs incurred. Fig. 8 represents the number of allocated blocks beyond the sum of the demand. What can be noticed is that two local minima, the first for low values and the second for values between 20 and 25. This behavior is the same for the two configuration databases with and without migration. However, when migration is used, between 3 and 4 PRBs on average can be saved. Furthermore, knowing that the sum of the demand for the 100 vectors is 360 PRBs, overallocation can be evaluated with migration at as low as 3%, which corresponds to parameter Ω. This low rate of overallocation gives a first proof of the effectiveness of our configuration-based allocation approach.

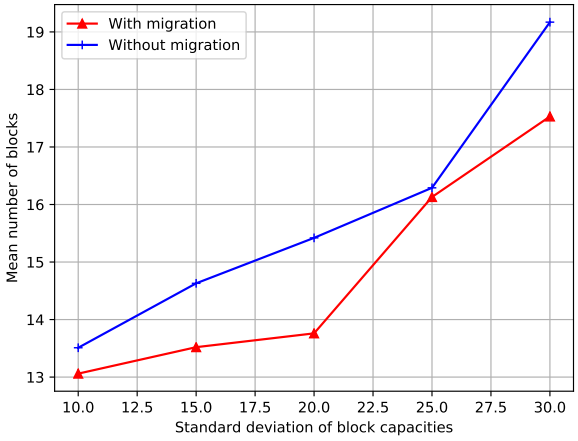

Fig. 9. Mean number of blocks generated by CBA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3074111, IEEE Internet of Things Journal

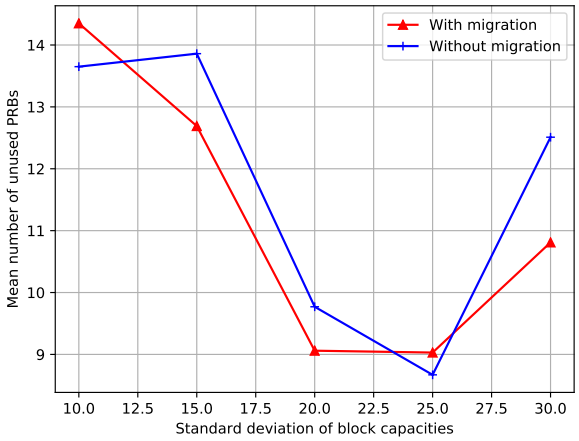

Fig. 10. Mean number of PRBs unused after the packing phase

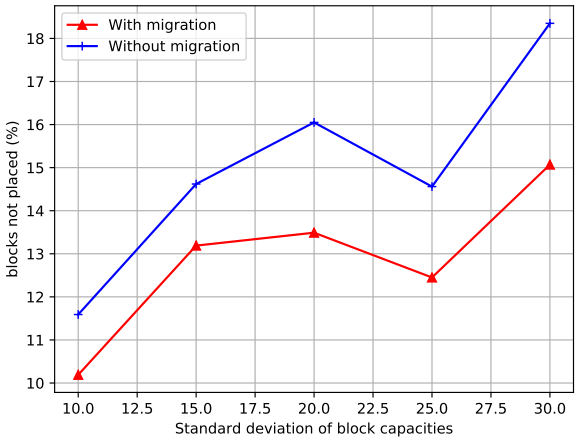

Fig. 11. Percentage of blocks used but not placed

The number of blocks incurred by the CBA is indicated in Fig. 9. It increases with the standard deviation. In fact, as the standard deviation increases, more small blocks are used. Therefore, if classes of TBs with a standard deviation of 25 are chosen, the number of blocks can be limited to around 16.

Let's continue our investigation with the packing phase. The number of unused PRBs in a rectangle of width 12 and height 30 is indicated in Fig. 10. Low and high standard deviation values are found to induce poor packing performance. In fact, with blocks of close capacity values, packing in the vertical direction might not be effective because it is likely that no block can adapt to the remaining vertical space. On the other hand, with very disproportionate blocks of capacity values, it would not be possible to pack as many large blocks as necessary, leaving several unused columns horizontally. Therefore, a possible conclusion is that using average standard deviation values between 20 and 25 would provide the best performance. Finally, no significant difference is observed between performance with and without migration, which could favor the configuration database without migration to benefit from its small size and thus reduce the complexity of CBA.

The last performance measure used is the percentage of blocks not placed. As this measure is linked to the number of blocks which increases with the standard deviation, an upward trend is also observed here, as indicated in Fig. 11. However, a local minimum is identified for a standard deviation of 25 where the percentage of unplaced blocks is as low as 12.8% for the database with migration and 14.6% for the database without migration.

Based on the four measures used in this study, what can be concluded is that a standard deviation of 25 would give the best performance for both databases with and without migration. Recall that the choice of the standard deviation is only an example to expose the different metrics likely to guide a good choice of profiles. Other parameters also deserve to be studied such as the average of the capacities.

TABLE V
COMPUTATION TIME IN SEC OF THE QOS FLOW TO TB MAPPING PROBLEM

	350	360	370	380	390	400
F2BAP'	18.17	21.47	21.66	19.82	28.15	37.11
CBA with migration	0.106	0.248	0.178	0.143	0.143	0.139
CBA without migration	0.054	0.054	0.055	0.047	0.054	0.049

D. Computation time

In addition to the efficient use of resources, real-time decision making is one of the most important objectives of our solution. In this study, different demand vectors have been created whose sum varies between 350 and 400 PRBs. The performance of the suggested configuration-based solution has been compared to the optimal solution provided by solving the F2BAP' problem (22). The resolution time of the assignment problem is reported in Table V. For the optimal solution, an average resolution time of 20 s is obtained, which is obviously not suitable for real-time applications. However, very low computation times are noticed when using our configuration-based solution. In fact, when the database with migration is used, an average computation time of 160 ms is obtained, but without migration, an average time of only 52 ms was recorded.

Finally, to check the quality of the solution found in each case, the average values of the objective-function are represented in Fig. 12. It is worth remarking that the solution provided by the configuration-based approach with migration is very close to the optimal solution, which indicates the precision of our configuration-based approach. However, the solution without migration uses on average 6 PRBs above the optimal solution.

E. Optimality gap

By this final study, an attempt is made at evaluating how close the CBAP solution is to the optimal solution. The relative gap will be used which measures the difference between the two solutions to the optimal. Note that our CBAP solution is

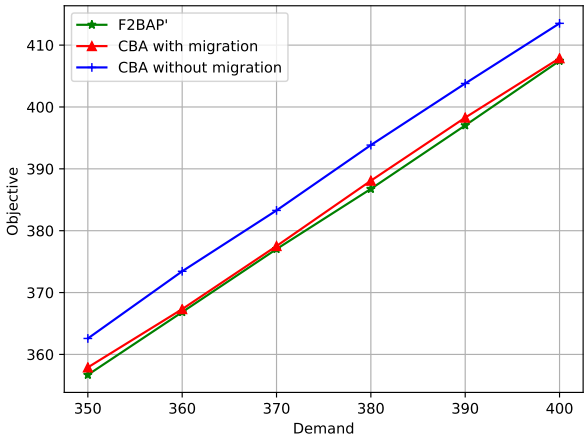

Fig. 12. Average value of the solution of the assignment problem

TABLE VI
UPPER BOUND ON THE OPTIMALITY GAP FOR THE CBAP SOLUTION

Total demand (PRB)	360	370	380	390	400
CBAP solution	350.72	351.46	351.45	349.79	350.29
Optimality gap (%)	2.58	2.37	2.38	2.84	2.7

given by sequentially resolving the assignment then the packing problems as shown in the flowchart in Fig. 4. However, the exact solution given by directly resolving the problem in section III-A or the Lagrangian duality in section III-B must result from an iterative resolution which would take a much longer computation time. Fortunately, the upper bound on the optimal solution is already known which corresponds to the maximum capacity of the grid, i.e. 360 PRBs. Therefore, there is no need to compute the optimal solution and only an upper bound on the optimality gap will be set which is computed using the following formula: $\frac{360-\text{CBAP solution}}{360}$. The profiles described by vector V_D in Table IV are used as well as 100 demand vectors whose sum varies between 360 and 400. The average results are summarized in Table VI.

The excellent performance of the proposed CBAP algorithm can easily be observed since the provided solution is at worst no further than 2.84% of the upper optimality bound.

VI. CONCLUSION

In this paper, a joint scheduling and packing algorithm was proposed, namely CBAP, that provided efficient and dynamic resource allocation for heterogeneous services such as eMBB and URLLC in 5G. Unlike most state-of-the-art algorithms which use the puncturing technique to allocate URLLC flows with the already used resources of the ongoing eMBB sessions and then try to reduce the performance damage to the interrupted eMBB sequences, what is proposed in this study is a dynamic allocation approach which sought a joint allocation of these different flows to blocks of resources from a database of configurations built in advance. Furthermore, the suggested approach considered the packing optimization problem of these blocks of resources inside the flexible 2D rectangular 5G frame. By carefully defining TB classes appropriate to the various parameters that describe the physical resources, CBAP achieved near-optimal allocations within a reasonable time frame. As a next step, it would be interesting to investigate the different optimization techniques based on ML to solve the online part of the CBAP and to quantify the possible gains in terms of processing time and quality of the solution.

ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to Prof. Ali M. AMRI of ENET'Com, University of Sfax, for his meticulous proofreading of this paper.

REFERENCES

[1] L. D. Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, Nov 2014.
[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, Fourthquarter 2015.
[3] S. S. Reka and T. Dragicevic, "Future effectual role of energy delivery: A comprehensive review of internet of things and smart grid," *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 90 – 108, 2018.
[4] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani, "Internet-of-things-based smart cities: Recent advances and challenges," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 16–24, Sep. 2017.
[5] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.
[6] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
[7] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Network*, vol. 32, no. 2, pp. 8–15, March 2018.
[8] 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, March 2016.
[9] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.
[10] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems," in *2016 IEEE Globecom Workshops (GC Wkshps)*, Dec 2016, pp. 1–7.
[11] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource Optimization With Flexible Numerology and Frame Structure for Heterogeneous Services," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2579–2582, Dec 2018.
[12] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, and P. Zhang, "Machine learning based flexible transmission time interval scheduling for embb and urllc coexistence scenario," *IEEE Access*, vol. 7, pp. 65 811–65 820, 2019.
[13] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1508–1522, 2017.
[14] E. Coffman, M. Garey, and D. Johnson, "Bin packing with divisible item sizes," *Journal of Complexity*, vol. 3, no. 4, pp. 406 – 428, 1987.
[15] E. G. Coffman, C. Courcoubetis, M. R. Garey, D. S. Johnson, P. W. Shor, R. R. Weber, and M. Yannakakis, "Perfect packing theorems and the average-case behavior of optimal and online bin packing," *SIAM Review*, vol. 44, no. 1, pp. 95–108, 2002.

[16] 3GPP, "5G; System architecture for the 5G System (5GS) ," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.501, 09 2020, 16.6.0 Release 16.

[17] S. Ahmadi, *5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*, 1st ed. Academic Press, June 2019.

[18] S. Manap, K. Dimyati, M. N. Hindia, M. S. Abu Talip, and R. Tafazolli, "Survey of radio resource management in 5g heterogeneous networks," *IEEE Access*, vol. 8, pp. 131 202–131 223, 2020.

[19] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 477–490, 2020.

[20] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC Traffic Within eMBB Services in 5G NR: Fair Scheduling," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1080–1093, 2021.

[21] W. Ning, Y. Wang, M. Liu, Y. Chen, and X. Wang, "Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services," *IEEE Access*, vol. 9, pp. 21 870–21 880, 2021.

[22] N. Ben Khalifa, V. Angilella, M. Assaad, and M. Debbah, "Low-complexity channel allocation scheme for urllc traffic," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 194–206, 2021.

[23] F. Librino and P. Santi, "Resource allocation and sharing in urllc for iot applications using shareability graphs," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10 511–10 526, 2020.

[24] Y. Huang, S. Li, C. Li, Y. T. Hou, and W. Lou, "A deep-reinforcement-learning-based approach to dynamic embb/urllc multiplexing in 5g nr," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6439–6456, 2020.

[25] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021.

[26] W. Sui, X. Chen, S. Zhang, Z. Jiang, and S. Xu, "Energy-efficient resource allocation with flexible frame structure for hybrid eMBB and URLLC services," *IEEE Transactions on Green Communications and Networking*, 2020.

[27] Y. Bengio, A. Lodi, and A. Prouvost, "Machine learning for combinatorial optimization: A methodological tour d'horizon," *European Journal of Operational Research*, vol. 290, no. 2, pp. 405 – 421, 2021.

[28] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 64–71, June 2017.

[29] 3GPP, "5G; NR; Physical channels and modulation," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211, 09 2020, 16.3.0 Release 16.

[30] ——, "5G; NR; NR and NG-RAN Overall description; Stage-2," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.300, 10 2020, 16.3.0 Release 16.

[31] C. Chang and N. Nikaein, "RAN Runtime Slicing System for Flexible and Dynamic Service Execution Environment," *IEEE Access*, vol. 6, pp. 34 018–34 042, 2018.

[32] G. Wäscher, H. Haußner, and H. Schumann, "An improved typology of cutting and packing problems," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1109 – 1130, 2007.

[33] A. Lodi, S. Martello, and M. Monaci, "Two-dimensional packing problems: A survey," *European Journal of Operational Research*, vol. 141, no. 2, pp. 241 – 252, 2002.

[34] D. Pisinger and M. Sigurd, "The two-dimensional bin packing problem with variable bin sizes and costs," *Discrete Optimization*, vol. 2, no. 2, pp. 154 – 167, 2005.

[35] 3GPP, "5G; NR; User Equipment (UE) radio access capabilities," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.306, 10 2020, 16.2.0 Release 16.

[36] Zukang Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, 2005.

[37] Inhyoung Kim, In-Soon Park, and Y. H. Lee, "Use of linear programming for dynamic subcarrier and bit allocation in multiuser OFDM," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 4, pp. 1195–1207, 2006.

[38] S. Chieochan and E. Hossain, "Adaptive radio resource allocation in ofdma systems: a survey of the state-of-the-art approaches," *Wireless Communications and Mobile Computing*, vol. 9, no. 4, pp. 513–527, 2009.

[39] M. L. Fisher, "The lagrangian relaxation method for solving integer programming problems," *Management Science*, vol. 27, no. 1, pp. 1–18, 1981.

[40] C. So-In, R. Jain, and A.-K. Al Tamimi, "eOCSA: an algorithm for burst mapping with strict QoS requirements in IEEE 802.16 e mobile WiMAX networks," in *Wireless Days (WD), 2009 2nd IFIP*. IEEE, 2009, pp. 1–5.

[41] H. W. Lenstra, "Integer programming with a fixed number of variables," *Mathematics of Operations Research*, vol. 8, no. 4, pp. 538–548, 1983.

[42] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Improved video support for Packet Switched Streaming (PSS) and Multimedia Broadcast/Multicast Service (MBMS) Services ," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 26.903, 07 2020, 16.0.0 Release 16.

[43] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5g usage scenarios and traffic models," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[44] Y. Boujelben, "Configuration-based assignment and packing in 5G," IEEE Dataport, 2020. [Online]. Available: http://dx.doi.org/10.21227/j8gr-hx72