IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 3, 2022, accepted March 18, 2022, date of publication March 30, 2022, date of current version April 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3163388

Scheduling of Heterogeneous Services by Resolving Conflicts

SOTIRIS SKAPERAS[1], (Member, IEEE), NASIM FERDOSIAN[2], (Member, IEEE), ARSENIA CHORTI[1], (Senior Member, IEEE), AND LEFTERIS MAMATAS[3], (Member, IEEE)

[1]ETIS UMR 8051, CY Cergy Paris University, ENSEA, CNRS, 95000 Cergy, France
[2]School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, WA 6102, Australia
[3]Department of Applied Informatics, University of Macedonia, 546 36 Thessaloniki, Greece

Corresponding author: Sotiris Skaperas (sotiris.skaperas@ensea.fr)

The work of Sotiris Skaperas, Nasim Ferdosian, and Arsenia Chorti was supported by the INEX Project eNiGMA of the Initiative of Excellence of CY University.

ABSTRACT Fifth generation (5G) new radio introduced flexible numerology to accommodate heterogeneous services. However, optimizing the scheduling of heterogeneous services with differing delay and throughput requirements over 5G new radio is a challenging task. In this paper, we investigate near optimal, low complexity scheduling of radio resources for ultra-reliable low-latency communications (URLLC) when coexisting with enhanced mobile broadband (eMBB) services. We demonstrate that maximizing the sum throughput of eMBB services while servicing URLLC users over a fixed length time-frequency grid, is, in the long-term, equivalent to minimizing the number of URLLC placements in the time-frequency grid; this is the consequence of reducing the number of infeasible placements for eMBB, to which we refer to as ''conflicts.'' To meet this new objective, we evaluate the performance of new, conflict-aware heuristics, consisting of a family of ''greedy'' and a lightweight heuristic inspired by bin packing optimization, all of near optimal performance. Moreover, having shed light on the impact of conflict in layer-2 scheduling, we investigate non-orthogonal multiple access (NOMA) as a potential approach for conflict resolution leveraging superposition coding. The superior performance of NOMA with respect to OMA, thanks to resolving conflicts, is showcased by extensive numerical results.

INDEX TERMS Time-frequency resource allocation, URLLC traffic, wireless communication, scheduling.

I. INTRODUCTION

The International telecommunication union (ITU) has defined new requirements and capabilities on 5G mobile communication systems to support a wide variety of new devices and services with diverse quality of service (QoS) requirements and characteristics [3]. The 3rd generation partnership project (3GPP) standardized 5G in the form of a novel radio interface technology, referred to as new radio (NR) [4]. 5G NR introduced flexible numerology and frame structure to accommodate heterogeneous service requirements, by supporting various values of subcarrier spacing and symbol / frame duration. Optimizing resource allocation in the NR numerology setting to deliver heterogeneous QoS requirements remains a challenging task [5]–[9].

In 5G and beyond, ultra-reliable low-latency communication (URLLC) services with extreme delay constraints will coexist with enhanced mobile broadband (eMBB) [10], that require very high bit rates (Gigabits per second) and have moderate latency (a few milliseconds) requirements [11]. Moreover, at present, URLLC services are expected to have lower traffic volumes than eMBB services [12], but this will not hold in the future for applications such as virtual reality and haptics. In this framework, the design of radio resource allocation strategies for URLLC traffic, when coexisting with eMBB, has been a focal point of recent research efforts [13]–[16].

In this direction, two approaches have been adopted by the 3GPP. The first is based on a ''puncturing'' framework: according to this, eMBB traffic is scheduled initially at the beginning of the slots; upon arrival of URLLC traffic, the latter is being prioritized and dynamically overlapped at mini-slots of ongoing eMBB transmissions (which are punctured, i.e., dropped). In the second approach, known as preemptive scheduling, resources are preemptively reserved for URLLC, before the demands are placed [1], [17], [18].

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

Based on puncturing scheduling, the studies in [8], [19]–[22] considered resource allocation strategies for the coexistence of URLLC and eMBB. The authors in [8] consider three types of models - threshold, linear and convex - to describe the eMBB data rate loss associated with the incoming URLLC traffic. Furthermore the authors in [20] propose a punctured scheduling approach for transmission of low latency communication traffic multiplexed on a shared channel with eMBB. Another approach is proposed in [21], where a risk-sensitive model was introduced in order to ensure URLLC allocation but also to minimize losses for eMBB users. However, these strategies can result in significant losses in terms of data rates for eMBB services and may also impact eMBB transmission reliability [23]. Targeting the above problem, the authors of [22] proposed a scheduling approach to maximize the minimum expected achieved rates and fairness, accounting for the expected values of the traffic. Alternatively, the authors in [1] studied the resource allocation of eMBB and URLLC services by preemptively reserving resources for URLLC. Such solutions ensure advantageous conditions for URLLC packets when they are generated, at the cost of wasting resources in absence of URLLC transmissions [9].

A flexible numerology and frame structure was explicitly considered in [1] by defining a time-frequency resource grid, containing several types of resource blocks of different shapes, expanding over different time spans and frequency ranges. Exploiting this flexibility to optimize the resource allocation to different services while ensuring their QoS requirements, was shown to be an *NP*-hard problem. The resource allocation optimization over flexible numerology and frame structure while avoiding the assignment of overlapping blocks that will cause collision (i.e., puncturing), still remains a challenging task.

In this paper, following the works in [1] and [8], we consider a 2-dimensional grid of resource blocks with different sizes in the time and frequency domains. The problem of identifying the resource allocation that maximizes the eMBB sum-rate is studied under the constraint of serving all URLLC throughput demands under different latency constraints ranging from 0.5 to 2 milliseconds (msec). In our previous work [24] we introduced a bin packing approach that minimizes the placements of the resource blocks for the URLLC services in order to minimize the infeasible placements for the eMBB services. In this work, additionally, we extend the above framework by proposing a second approach based on low complexity algorithms, that manage the infeasible placements. We also extend our previous bin packing approach incorporating a preprocessing step that counts the overall feasibility of the grid. We note that in our formulation, unlike in [22] and other published work, we do not account for the expected traffic but rather for the actual traffic in an appropriate time-frequency grid. This approach has also been followed in [1] among others.

Furthermore, we consider and formulate an alternative non-orthogonal multiple access (NOMA) scheduling proposal. In particular, we introduce a 2-user NOMA scheme based on pairing eMBB and URRL services at the minislot level and show that the proposed approach achieves higher eMBB sum rates when compared to the optimal orthogonal scheduling as it avoids puncturing and preemptive allocation.[1] Finally, we evaluate the performance of the proposed algorithms under several numerologies (fixed, flexible and multiple). More precisely, the main contributions of this work are outlined below:

1) We first re-formulate the problem of eMBB throughput maximization, introducing the URLLC conflicts minimization in the objective function. The novel concept of ''conflict'' captures the penalties occurring due to the fact that orthogonal multiple access (OMA) does not allow overlapping of resources; as a result, OMA scheduling incurs a large number of infeasible resource allocation combinations. To the best of our knowledge, our earlier conference paper [24] is the first in the literature introducing conflict-aware solutions for the problem at hand.
2) Next, we propose three conflict-aware, multi numerology radio resource allocation heuristics to maximize scheduling efficiency for URLLC, when coexisting with eMBB services. Three different functions of the i) average, ii) the instantaneous (placement specific), or iii) the aggregate conflict are used to normalize the throughput utility function and incorporate penalties, when increasing conflicts. We argue and showcase through extensive simulation results that employing the proposed utilities improves the performance of proposed algorithms in the literature, as this in [1].
3) Subsequently, we depart on a completely different approach with a high accuracy and low computational complexity. We treat the scheduling problem as a specific instance of bin packing optimization, solved by minimizing the placements of URLLC services in the time-frequency resource grid; to this end, we propose to group the resource blocks in different categories with respect to URLLC demands. Within each category, we solve a knapsack maximization of the sum eMBB throughput. Our proposal builds on previous results in [24], [25] and is inspired by the refined-first-fit family of heuristics to solve bin packing problems. Simulation results show that the novel heuristic algorithm, of complexity $N \log(N)$, provides a lightweight and near optimal solution to the resource allocation scheduling of URLLC, when coexisting with eMBB.
4) Furthermore, having clarified the importance of minimizing conflicts between different services, the utilization of NOMA schemes [26], [27] provide as a competitive candidate for interference management. NOMA allows the superposition of services, even at the mini-slot level by employing superposition coding at

[1] In this work we focus on 2-user NOMA as opposed to multi-user NOMA as the latter has been shown to be vulnerable to error cascading.

the transmitter and successive interference cancellation at the receivers [28], [29]. Although most works on NOMA utilize the aspect of increased spectral efficiency to showcase superiority with respect to OMA, we further provide strong motivation for adopting NOMA as a conflict mitigation approach in scheduling problems. An extensive set of numerical results, investigating 2-user and multiple-user NOMA's performance, for both fixed and flexible numerology, shows the significant gains in terms of sum eMBB throughput, when adopting NOMA in a flexible numerology setting.

The rest of this paper is organized as follows: Section II presents the resource allocation optimization problem along with an equivalent formulation as a conflict minimization problem. Conflict-aware heuristic algorithms are described in Section III, while the problem re-formulation when using NOMA is presented in Section IV. Section V presents numerical results showing the near-optimal performance of the proposed heuristics as well as the superiority of NOMA for URRLC and eMBB coexistence, both in the case of flexible as well as fixed numerologies. Finally, conclusions are given in Section VI.

II. PROBLEM FORMULATION

We first provide a review of basic concepts in 5G NR flexible numerology and detail the considered scheduling problem.

A. BACKGROUND ON 5G NR FLEXIBLE NUMEROLOGY

5G NR Release-15 [4] defines a flexible numerology with subcarrier spacing (SCS) of 15, 30, and 60 kHz below 6 GHz, and 60 and 120 kHz above 6 GHz, compared to long-term evolution (LTE) which uses a fixed numerology with SCS of 15 kHz below 6 GHz. 5G NR also defines a 10 msec frame, with each frame divided into 10 subframes of 1 msec, which are further divided into one or more mini-slots. A mini-slot comprises 14 OFDM symbols for a configuration using normal cyclic prefix, or 12 OFDM symbols for extended cyclic prefix.

In 5G NR, the mini-slot size is defined according to the symbol duration, which is inverse to the SCS, to ensure the orthogonality of the subcarriers. By using higher SCS, the symbol duration decreases and hence also the mini-slot size, which is beneficial for lower latency [10]. URLLC traffic requires extremely low delays, often lower than 1 ms [30]. The URLLC latency requirements can only be satisfied if the transmission duration and round-trip-time (RTT) are shorter than the corresponding latency constraint.

B. SCHEDULING PROBLEM FORMULATION

We focus, in this work, on downlink scheduling, similarly to the majority of the existing related works in the literature, e.g., [13]–[23]. The system model and the scheduling framework follow the structure of [1], [8]. Time is divided into slots, with 2 msec slot duration. Each slot serves both throughput hungry (eMMB) and ultra-low latency users (URLLC), which have to be serviced until the next slot. In addition, the latter have also to satisfy specific latency requirements. Moreover, we assume that URLLC arrivals follow a Poisson distribution with parameter λ, while two cases are considered for the eMBB services: i) full buffer model, and, ii) $|\mathcal{K}^{(c)}| = |\mathcal{K}| - |\mathcal{K}^{(\ell)}|$. The objective is to find the resource allocation in each slot that maximizes the sum throughput of the eMBB while satisfying the throughput demands and latency constraints of the latter. Finally, we utilized [2] as a tool to implement the time-frequency grid.

The terminology employed in the rest of the paper is tabulated in Table 1: $\mathcal{K}$ denotes the set of all services, $\mathcal{K}^{(c)}$ the set of eMBB users, $\mathcal{K}^{(\ell)}$ the set of URLLC users, $\mathcal{B}$ is the set of all possible resource blocks according to the numerology employed and $\mathcal{I}$ denotes the set of all mini-slots. Moreover, the parameter q_k denotes the throughput demand for the

TABLE 1. Notation table.

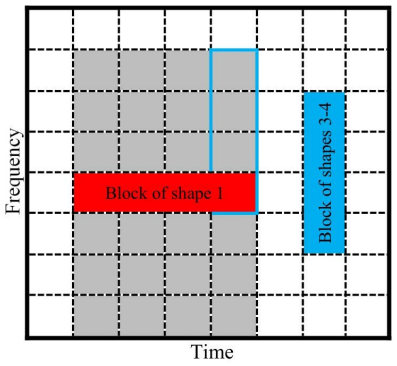
Notation	Sets
$\mathcal{K}^{(\ell)}$	set of URLLC services
$\mathcal{K}^{(c)}$	set of eMBB services
$\mathcal{K}$	set of all services, for which $\mathcal{K} = \mathcal{K}^{(\ell)} \cup \mathcal{K}^{(c)}$
$\mathcal{B}$	set of candidate blocks with respect to the numerology
$\mathcal{I}$	set of all basic units of the grid, i.e., the minimum unit of resource in the time-frequency domain (mini-slots)
Notation	**Parameters**
τ_k	maximum delay tolerance of service $k \in \mathcal{K}$
q_k	throughput demand (in bits) of service $k \in \mathcal{K}^{(\ell)}$
$r_{b,k}$	throughput of each resource block $b \in \mathcal{B}$ assigned to the service $k \in \mathcal{K}$
t_b	end time of block $b \in \mathcal{B}$
$\alpha_{b,i}$	parameter representing if block $b \in \mathcal{B}$ includes minislot $i \in \mathcal{I}$
$c_{b,p}$	parameter representing if two blocks, $b, p \in \mathcal{B}$, are overlaping
$u_{b,k}$	utility of each resource block $b \in \mathcal{B}$ assigned to service $k \in \mathcal{K}$
N	total number of subcarriers
$SINR_{k,i,n}$	signal to interference and noise ratio of minislot $i \in \mathcal{I}$, of subcarrier $n \in N$, assigned to service $k \in \mathcal{K}$
$C_{k,i}$	achievable Shannon rate of each minislot $i \in \mathcal{I}$, assigned to service $k \in \mathcal{K}$
η_j	symbol duration of the j block shape
TTI_j	transmission time interval duration of the j block shape
SCS_j	subcarrier spacing of block shape j.

$k \in \mathcal{K}^{(\ell)}$ services that has to be satisfied with a strict latency tolerance requirement of τ_k. Additionally, we introduce the binary parameter $\alpha_{b,i}$, $b \in \mathcal{B}$, $i \in \mathcal{I}$ which indicates whether a block $b \in \mathcal{B}$ includes basic unit $i \in \mathcal{I}$, in which case $\alpha_{b,i} = 1$, otherwise $\alpha_{b,i} = 0$.

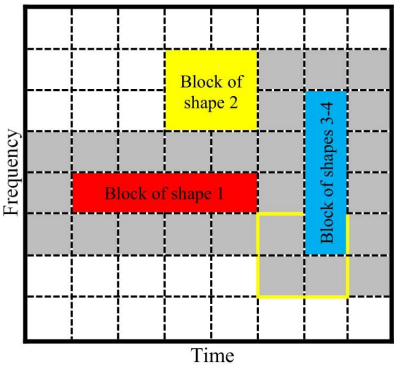
In Table 2 we describe the most widely utilized resource block specifications for 5G NR, depicted in Fig. 1(a) and (b); resource blocks of shape 1 shown in red, resource blocks of shape 2 shown in yellow and resource blocks of shapes 3−4 shown in blue (employing flexible numerology, $\mathcal{K}^{(c)}$ and $\mathcal{K}^{(\ell)}$ can utilize any of the given shapes). To demonstrate the concept of conflict, in Fig. 1(a), we illustrate in gray shade the invalid placements for shapes 3-4 when a specific placement of shape 1 has taken place, while in Fig. 1(b) we show the invalid placements for blocks of shape 2, when an additional placement of shape 3-4 has been decided.

TABLE 2. Feasible resource blocks options in flexible numerology, according to [4].

	Shape 1	Shape 2	Shape 3	Shape 4
TTI duration (ms)	0.5	0.25	0.125	0.125
SCS (kHz)	15	30	60	60
Symbol duration (μs)	66.7	33.3	16.7	16.7
CP (μs)	4.7	2.3	1.2	4.17
Number of Symbols	7	7	7	6



(a) Conflicts (in grey) with respect to a block of shapes 3-4, when a block of shape 1 has already been placed. The blue line indicates a conflicted block of shapes 3-4.



(b) Conflicts (in grey) with respect to a block of shape 2, when a block of shape 1 and a block of shapes 3-4 has already been placed. The yellow line indicates a conflicted block.

FIGURE 1. Resource allocation in a (time) slot, considering the flexible numerology context, with three types of resource blocks and the corresponding conflicts (grey).

Now, we can define the achievable throughput of each block $b \in \mathcal{B}$ assigned to service $k \in \mathcal{K}$, denoted by $r_{b,k}$, which depends on the signal to interference and noise ratio (SINR) and the configuration of the block (including the parameters in Table 2). More precisely, we first define the achievable Shannon rate of each minislot $i \in \mathcal{I}$ assigned to service $k \in \mathcal{K}$, as follows,

$$C_{k,i} = \sum_{n=1}^{N} \log_2(1 + SINR_{k,i,n}), \quad i \in \mathcal{I},\ k \in \mathcal{K}, \tag{1}$$

where N is the total number of subcarriers. Accounting for the impact of the cyclic prefix, the rate per minislot is given by,

$$E_{k,i} = \eta_j C_{k,i}, \quad j = \{1, 2, 3, 4\}, \tag{2}$$

where $\eta_j = \frac{T_j}{T_j + T_{cp}}$, T_j is the symbol duration of the $j \in \{1, 2, 3, 4\}$ block shape and T_{cp} is the cyclic prefix (CP) length. Then, the achievable throughput of each resource block $b \in \mathcal{B}$ with respect to service $k \in \mathcal{K}$ can be expressed as follows,

$$r_{b,k} = \sum_{i=1}^{\mathcal{I}} (a_{b,i} E_{k,i}) TTI_j SCS_j \times \mathbf{1}_{\{\tau_k - t_b > 0\}}, \quad j = \{1, 2, 3, 4\}, \tag{3}$$

where TTI_j and SCS_j are the transmission time interval duration and the subcarrier spacing of the $j \in \{1, 2, 3, 4\}$ block shape, respectively. Additionally, $\mathbf{1}_{\{x\}}$ is the indicator function for the logical proposition x. Note that the delay constraint is incorporated in the problem by considering that the end time t_b of the block $b \in \mathcal{B}$ has to comply with the delay tolerance τ_k of the service $k \in \mathcal{K}$, otherwise the specific block is infeasible for the specific service. As such, the latency tolerance constraints of the URLLC services need not to appear explicitely in the problem formulation presented in the following. In the rest of the paper, by $x_{b,k}$ we denote a binary variable that takes the value 1 if the resource block $b \in \mathcal{B}$ is assigned to service k, otherwise $x_{b,k} = 0$.

A common objective in eMBB and URLLC coexistence is articulated in maximizing the sum throughput of $\mathcal{K}^{(c)}$ services under the constraint of satisfying the latency and throughput demands of $\mathcal{K}^{(\ell)}$, without any overlapping between the allocated resource blocks. In other words, our goal is to find the resource allocation that satisfies the URLLC users' demands, with minimal losses for eMBB users in terms of throughput, and, subsequently schedule all the remaining resource blocks

to the eMBB services. The general problem formulation is given as follows:

$$[\text{P0}] \quad \max_{x_{b,k}\in\{0,1\}} \sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}^{(c)}} r_{b,k}x_{b,k}, \tag{4}$$

$$\text{s.t.} \sum_{b\in\mathcal{B}} r_{b,k}x_{b,k} \geq q_k, \quad k \in \mathcal{K}^{(\ell)}, \tag{5}$$

$$\sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}} a_{b,i}x_{b,k} \leq 1, \quad i \in \mathcal{I}. \tag{6}$$

In [1] it was proven that the combinatorial problem P0 is an *NP*-hard partition problem and a heuristic algorithm was proposed, referred to, in the rest of the paper, as the *baseline heuristic*, which uses a utility matrix $\mathbf{u}$ with elements $u_{b,k}$ that represent the utility of a block $b \in \mathcal{B}$ assigned to a specific service $k \in \mathcal{K}$. Then, in the first step of the heuristic algorithm, the block b is allocated to service $k \in \mathcal{K}^{(\ell)}$ with the maximum $u_{b,k}$; notice that choosing the allocation that maximizes the utility without at the same time examining the ''cost'' of this placement in terms of generated conflict is clearly sub-optimal. The step is iterated until all the demands for $k \in \mathcal{K}^{(\ell)}$ are satisfied under the constraint (5). Next, in the second step, the placements for $k \in \mathcal{K}^{(c)}$ services are allocated, using a similar principle, until no other non-overlapping blocks have remained. Hence, the placement of the $\mathcal{K}^{(\ell)}$ and $\mathcal{K}^{(c)}$ has been treated as two separate resource allocation problems. The complexity of the baseline heuristic algorithm was shown to be $\mathcal{O}(|\mathcal{B}||\mathcal{K}|\log(|\mathcal{B}||\mathcal{K}|))$, without accounting for the computation of utility matrices.

The baseline heuristic has been extended in [1] to incorporate other utility matrices denoted by $\mathbf{u}_{LP}, \mathbf{u}_{LD} \in \mathbb{R}_{\mathcal{B}\times\mathcal{K}}$, where $\mathbf{u}_{LP}$ and $\mathbf{u}_{LD}$ denote the optimal solutions of the linear programming (LP) and the Lagrange dual (LD) relaxation of P0, respectively. With these two new utilities, an extension of the baseline heuristic was proposed to calculate concurrently the solution of the heuristic algorithm by adopting both $\mathbf{u}_{LP}$ and $\mathbf{u}_{LD}$ utilities and retaining the best result between them; this allowed to reach a near-optimal performance, at the cost of high computational complexity, especially considering that the dual problem P0-LD also applies a sub-gradient method.

Discussing the above approach, whose basic principle (with few variations) can be found in other published work, e.g., [8], we notice that despite the fact that the overall aim is to *jointly* maximize the throughput of $\mathcal{K}^{(c)}$ while meeting the demands of $\mathcal{K}^{(\ell)}$ services, these two interwoven goals are treated separately; in order to satisfy constraint (5), *first* the demands of URLLC services are met and *then* the placements of eMBB services take place.

Such policies solve P0 by accounting only for constraint (5), which is suboptimal as they do not consider the impact of the $\mathcal{K}^{(\ell)}$ services allocation to the consequent allocation of the $\mathcal{K}^{(c)}$ services, i.e., constraint (6). We notice that previously proposed algorithms operate on a single optimization target at any instance, that of maximizing first the URLLC throughout and then maximizing the eMBB throughput. Building on this observation, we will first show that the previously presented baseline heuristic can be improved, if the conflict is taken explicitly into account.

To this end, we introduce an explicit description of the impact that the assignment of any resource block to a specific service has on the feasible assignments of the remaining blocks. In other words, we account for the amount of generated conflict by any specific URLLC or eMBB resource block placement. To evaluate the impact of constraint (6) explicitly, we define the conflict as

$$c_{b,p} = \begin{cases} 1, & \text{if } \sum b \in \mathcal{B} \sum_{p\in\mathcal{B}} (\alpha_{b,i} + \alpha_{p,i}) > 1, \\ & i \in \mathcal{I}, \quad b \neq p \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

for $b, p \in \mathcal{B}$. As a next step we note that,

$$\sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}^{(c)}} r_{b,k}x_{b,k} = R_{total} - \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{B}} \sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k}, \tag{8}$$

where R_{total} denotes the maximum sum throughput of the whole slot with respect to $\mathcal{K}^{(c)}$ and the second triple sum represents the losses in $\mathcal{K}^{(c)}$ throughput, because of the conflicts generated by the placements of all services. Given that R_{total} has a specific value (that can be explicitely evaluated) for any particular slot realization (based on the specific channel realizations), the maximization of (4) is equivalent (from an optimization point of view) to the minimization of the aggregate conflict, i.e.,

$$\max_{x_{b,k}\in\{0,1\}} \left(R_{total} - \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{B}} \sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k} \right) \Leftrightarrow \min_{x_{b,k}\in\{0,1\}} \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{B}} \sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k}. \tag{9}$$

Hence, the maximization of the sum eMBB throughput may be reduced to the minimization of the potential conflicts. We also note that:

$$\mathbb{E}\left[\sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{B}} \sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k} \right] = |\mathcal{C}|\,\bar{r}, \tag{10}$$

where $\mathbb{E}[\cdot]$ denotes expectation over the channel realizations, $\mathcal{C}$ is the set of conflicts when all resource blocks have the same average throughput $\bar{r} = \mathbb{E}\left[r_{b,k}\right]$ and $|\cdot|$ denotes cardinality; i.e., from (6) and (7) it emerges that *we need, on average, to minimize the number of conflicts.*

Considering these remarks, we propose novel heuristic algorithms for P0, focusing on minimizing the number of placements of $\mathcal{K}^{(\ell)}$ services. The first set of heuristics, dubbed in the following as conflict-aware greedy, use ''conflict'' enhanced variations of the utility proposed in the baseline heuristic and aim at closing the optimality gap. The second approach is built on an interpretation of (6) as a bin packing optimization problem [31]; based on this approach we develop a lightweight scheduling approach that is shown to be near-optimal.

Furthermore, as the minimization of conflicts is shown to be an equivalent optimization objective to the sum throughput maximization, we propose the use of NOMA to allow for overlapping of placements. The proposed heuristics and NOMA approaches are detailed in the next two sections.

III. HEURISTIC ALGORITHMS FOR CONFLICT RESOLUTION

A. CONFLICT-AWARE HEURISTIC SOLUTIONS

We first propose extensions of the baseline heuristic, in [1], [8], etc., by introducing penalties in URLLC resource allocations, expressed as functions of the conflict. To this end, we introduce two metrics for the conflict induced by $\mathcal{K}^{(\ell)}$ services allocation. The aggregate conflict C_b^t,

$$C_b^t = \sum_{p \in \mathcal{B}} c_{b,p}, \quad p, b \in \mathcal{B}, \tag{11}$$

that measures the total number of overlapping blocks with the block b, and, the average conflict $C_{b,k}^r$,

$$C_{b,k}^r = \sum_{p \in \mathcal{B}} \frac{c_{b,p} r_{p,k}}{C_b^t}, \quad p, b \in \mathcal{B} \text{ and } k \in \mathcal{K}^{(\ell)} \tag{12}$$

that corresponds to the average throughput – for every service $k \in \mathcal{K}^{(\ell)}$ – of the blocks $p \in \mathcal{B}$ that overlap with block $b \in \mathcal{B}$.

Using these new conflict measures, we propose three variations for the utility matrix to be used in solving P0:

- In the first version the utility becomes,

$$u_{b,k}^{total} = \frac{r_{b,k}}{C_b^t};$$

- In the second variation, the utility becomes,

$$u_{b,k}^{avg} = \frac{r_{b,k}}{C_{b,k}^r};$$

- Finally, in the third variation, we use the following utility,

$$u_{b,k}^{last\ pl.} = \begin{cases} r_{b,k}, & \text{if } k = 1, \ldots, \left|\mathcal{K}^{(\ell)}\right| - 1 \\ u_{b,k}^{avg}, & \text{if } k = \{\left|\mathcal{K}^{(\ell)}\right|\}. \end{cases}$$

The utility matrix $u_{b,k}^{last\ pl.}$ is introduced to incorporate a "compromise" between the baseline and the conflict-aware approaches; notably, it considers the impact of the conflict only in the last $\mathcal{K}^{(\ell)}$ service placement, since our simulations revealed that in this last placement, usually, more blocks are required to satisfy the demands constraint.

In Algorithm 1 we outline, in the form of pseudocode, the proposed conflict-aware heuristic. Algorithm consists of two phases: i) in Lines 1-10 we decribe the allocation of the URLLC services, and, ii) in Lines 11-14 the allocation of the eMBB services is described. First, in Line 2, we denote the set $\mathcal{G}$, consisting of the resource blocks to be allocated for the URLLC services (initially $\mathcal{G} = \emptyset$). Set $\mathcal{G}$ is augmented with the couple (b', k') that maximizies the conflict utility metric ($\mathbf{u}^{total}$, $\mathbf{u}^{avg}$ and $\mathbf{u}^{last\ pl.}$) across all the available $b \in \mathcal{B}$ and $k \in \mathcal{K}^{(\ell)}$, described in Line 3. A service

Algorithm 1 Conflict-Aware Resource Allocation Algorithm (*CA*) Based on [1]

Input: $\mathbf{u}^{(\ell)} = [u_{b,k}]$, $b \in \mathcal{B}, k \in \mathcal{K}^{(\ell)}$, utility matrix for $K^{(\ell)}$ ($\mathbf{u}^{total}$, $\mathbf{u}^{avg}$ or $\mathbf{u}^{last\ pl.}$) and $\mathbf{u}^{(c)} = [r_{b,k}]$, $b \in \mathcal{B}, k \in \mathcal{K}^{(c)}$, utility matrix for $K^{(c)}$.

Output: Block-service assignment $\mathcal{G}$.

► **Phase ($\mathcal{K}^{(\ell)}$ resource allocation):**

1: **repeat**
2: Remove from $\mathcal{B}$ the blocks in $\mathcal{G}$ and the overlapping with $\mathcal{G}$ blocks.
3: $(b', k') \leftarrow \text{argmax}_{b \in \mathcal{B}, k \in \mathcal{K}^{(\ell)}} u_{b,k}^{(\ell)}$, $\mathcal{G} \leftarrow \mathcal{G} \cup \{(b', k')\}$.
4: **if** $q_{k'}$ is met **then**
5: $\mathcal{K}^{(\ell)} \leftarrow \mathcal{K}^{(\ell)} \setminus k'$.
6: **end if**
7: **until** $\mathcal{K}^{(\ell)} = \varnothing$ **or** $\mathcal{B} = \varnothing$
8: **if** $\mathcal{K}^{(\ell)} \neq \varnothing$ **then**
9: The demand of the remaining users in $\mathcal{K}^{(\ell)}$ cannot be met.
10: **end if**

► **Phase ($\mathcal{K}^{(c)}$ resource allocation):**

11: **repeat**
12: Remove from $\mathcal{B}$ the blocks in $\mathcal{G}$ and the overlapping with $\mathcal{G}$ blocks.
13: $(b', k') \leftarrow \text{argmax}_{b \in \mathcal{B}, k \in \mathcal{K}^{(c)}} u_{b,k}^{(c)}$, $\mathcal{G} \leftarrow \mathcal{G} \cup \{(b', k')\}$.
14: **until** $\mathcal{B} = \varnothing$

$k \in \mathcal{K}^{(\ell)}$ is satisfied when its q_k demand is met and the first phase is concluded when all services are satisfied or no other available $b \in \mathcal{B}$ exist, according to Lines 4-6 and Lines 7-10, respectively.

In case all URLLC services are satisfied, algorithm proceeds to the second phase. From the available blocks $\mathcal{B}$ we exclude the resource blocks used for the URLLC services and all the overlapping to these blocks (Line 12). Finally, the eMBB services are allocated iteratively, according to the utility metric $u_{b,k}^{(c)} = r_{b,k}$ (Line 13), until $\mathcal{B} = \emptyset$ (Line 14).

B. HEURISTIC INSPIRED FROM BIN PACKING OPTIMIZATION

In the standard bin packing problem formulation, the goal is to find the optimal placement of items of different volumes in the minimum number of containers (bins) of fixed volume [31]. Although the bin packing is a combinatorial *NP*-hard problem, due to it's widespread encounter in a large number of settings, various proposed heuristics have been reported in the literature with different optimality gaps. Here, we propose a novel, computationally efficient scheduling approach, inspired by the refined-first-fit heuristic for the standard bin packing problem.

The proposed scheduling heuristic that accounts for conflicts is summarized in Algorithm 2, jointly minimizing the number of $\mathcal{K}^{(\ell)}$ resource allocations (placements) and throughput losses for $\mathcal{K}^{(c)}$ users. Allocation of resources to $\mathcal{K}^{(\ell)}$ and $\mathcal{K}^{(c)}$ services is treated sequentially but still in an

Algorithm 2 Bin Packing Based Resource Allocation Algorithm (BPB)

Input: throughput matrix $\mathbf{r} = [r_{b,k}], b \in \mathcal{B}, k \in \mathcal{K}$, aggregated-throughput-loss vector $\mathbf{e} = [e_b], b \in \mathcal{B}$ and overall number of categories (bins) H.
Output: Block-service assignment $\mathcal{G}$.

► **Phase (generate categories):**
1: **for** $k = 1$ to $|\mathcal{K}^{(\ell)}|$ **do**
2: **for** $i = 1$ to H **do**
3: Cat^iU^k contains all resource blocks $b \in \mathcal{B}$ where $\lceil q_k/r_{b,k} \rceil = i$ (14);
4: Check pairwise conflicts among categorized blocks and remove the blocks with the higher $\mathbf{e_b}$, $b \in \mathcal{B}$ (13);
5: **end for**
6: **end for**
► **Phase ($\mathcal{K}^{(\ell)}$ resource allocation):**
7: **for** $k = 1$ to H **do**
8: **repeat**
9: **if** $|Cat^iU^k| \geq i$ **then**
10: $B' \leftarrow \text{argsort}(\mathbf{e}_b), b \in Cat^iU^k$;
11: $\mathcal{G} \leftarrow \mathcal{G} \cup (B'(1,i), k)$,
12: Remove from $\mathcal{B}$ the blocks in $\mathcal{G}$ and those overlapping with the blocks in $\mathcal{G}$;
13: **end if**
14: **if** q_k is not met **then**
15: Demand of $k \in \mathcal{K}^{(\ell)}$ can be satisfied
16: **end if**
17: **until** q_k is met **or** $i = H$
18: **end for**
► **Phase ($\mathcal{K}^{(c)}$ resource allocation):**
19: **repeat**
20: Remove from $\mathcal{B}$ the blocks in $\mathcal{G}$ and those overlapping with the blocks in $\mathcal{G}$;
21: $(b', k') \leftarrow \arg\max_{b \in \mathcal{B}, k \in \mathcal{K}^{(c)}} r_{b,k}$, $\mathcal{G} \leftarrow \mathcal{G} \cup (b', k')$;
22: **until** $\mathcal{B} = \emptyset$

interwoven approach, with URLLC being served first to meet the latency requirements. In the following, we denote by vector $\mathbf{e}$ the aggregated throughput losses for each allocation of a block $b \in \mathcal{B}$, i.e.,

$$e_b = \sum_{p \in \mathcal{B}} \sum_{k \in \mathcal{K}^{(c)}} c_{b,p} r_{b,k}. \tag{13}$$

The proposed heuristic contains three steps: i)generation of bin-packing categories, in Lines 1-6, ii) URLLC resource allocation, in Lines 7-18, and, iii) eMBB allocation, Lines 19-22. First, for each $k \in \mathcal{K}^{(\ell)}$ we generate H categories (bins) with decreasing fractional sizes with respect to $q_k, k \in \mathcal{K}^{(\ell)}$. Category $i \in \{1, \ldots, H\}$ is defined as the set of all resource blocks $b \in \mathcal{B}$ for which the ceiling of the service demand ratio over the throughput of block b is equal to i, or equivalently, category Cat^iU^k contains the available resource blocks which satisfy at least $1/i$-th of the service demand q_k. Formally, we define,

$$Cat^iU^k = \left\{ b : \left\lceil \frac{q_k}{r_{b,k}} \right\rceil = i, \forall b \in \mathcal{B} \setminus \{Cat^jU^k\}_{j=1,\ldots,i-1} \right\}, \quad k \in \mathcal{K}^{(\ell)}, \quad i \in \{1, \ldots, H\}. \tag{14}$$

The above operation is described in Lines 1-3. Afterwards, in Line 4, we remove the overlapping blocks with the higher aggregated throughput loss (within each category), using (13).

In other words, Cat^1U^1 is the category of the blocks which individually satisfy the whole demand of the URLLC service $k = 1$. Therefore, the categories created for service $k \in \mathcal{K}^{(\ell)}$ range from Cat^1U^k – containing the most valuable blocks (valuable in terms of throughput $r_{b,k}$) – till Cat^HU^k, containing the least valuable blocks in order. Note that i) we need *at most* i elements from Cat^iU^k to satisfy the demand q_k of service $k \in \mathcal{K}^{(\ell)}$; ii) categories might be empty, so H needs to be defined according to the expected throughput per mini-slot, as well as its variance.

Next, we consider the allocation of the URLLC services. For each $k \in \mathcal{K}^{(\ell)}$ we select the first category Cat^iU^k with elements at least equal to $i \in H$, Lines 7-9. In this category, we subsequently introduce a further minimization problem in order to select the elements from each category that incur the minimum loss to eMBB, i.e.,

$$\min_{y_b \in \{0,1\}} \sum_b e_b y_b, \quad b \in (\mathcal{K}^{(l)} \cap Cat^iU^k)$$
$$\text{s.t.} \sum_{b \in Cat^iU^k} y_b \leq i. \tag{15}$$

Note that if (15) is interpreted as a knapsack problem, each element of a given category has the same weight (equal to unity), while the values (losses in the specific instance) differ. Similar problems are encountered in different settings, e.g., the subcarrier resource allocation in [25]. Exploiting these previous results, we reproduce a simple heuristic according to which the elements of each category are reordered[2] $\mathcal{O}\big(\max_{i,k}\{|Cat^iU^k| \log(|Cat^iU^k|)\}\big)$. in increasing aggregated loss $e_b, b \in \mathcal{B}$, Line 10. Subsequently, the first i elements of category Cat^iU^k are allocated to URLLC, Line 11. After each allocation, the allocated blocks are removed from $\mathcal{B}$ and all other categories, Line 12. The procedure is repeated until the demand $q_k, k \in \mathcal{K}^{(\ell)}$ is satisfied or no more categories exist for the specific service k, Line 17; in this last case solution is infeasible, Lines 14-16.

As an example, after this step, the first element of Cat^1U^k is the resource block that can simultaneously cover the demand q_k of URLLC service k while incurring the least aggregate losses for the eMBB users. The joint minimization of the number of $\mathcal{K}^{(\ell)}$ placements and the losses due to conflicts is achieved simply by assigning to service $k \in \mathcal{K}^{(\ell)}$ the first i elements of Cat^iU^k, starting from $i = 1$, i.e., the allocation for demand q_k starts from Cat^1U^k. As explained

[2]The ordering has a complexity.

before, the most valuable categories in terms of throughput satisfy URLLC services by using the least number of resource blocks and result in the minimum number of $\mathcal{K}^{(\ell)}$ placements, that is expected on average to incur the minimum losses due to conflicts. Furthermore, having re-ordered the elements of each category in increasing eMBB loss value, we jointly account for both constraints (5) and (6) in one go.

In the last phase of the algorithm, Lines 20-22, the resource allocation to $\mathcal{K}^{(c)}$ services takes place. This is performed by selecting the block-service pairs with the highest throughput $r_{b,k}, b \in \mathcal{B}, k \in \mathcal{K}^{(c)}$) from the remaining available blocks. The latter have not been allocated to a URLLC service, since once a block is allocated it is removed from $\mathcal{B}$. This step is iterated until no more blocks remain available.

Finally, we also consider a modified version of the bin packing based heuristic (mBP), targeting on challenging time-frequency grids, where infeasibility is the major issue. In this case, we introduce a pre-processing step to check the feasibility of each slot. We first count the total throughput of all available block placements and compare with throughput resulting from the placement of all the available blocks for the URLLC services, in both cases with respect to the constraint (3). Then if,

$$\sum_{b' \in \mathcal{B}} \sum_{k \in \mathcal{K}^{(\ell)}} r_{b',k} > \delta \sum_{b' \in \mathcal{B}} \sum_{k \in \mathcal{K}} r_{b',k},$$

where $b' \in \mathcal{B}$ are the blocks that satisfy constraint (3) and $\delta \in (0, 1)$, instead of using the e_b metric for the allocation of the $k \in \mathcal{K}^{(\ell)}$ services we switch the metric to $e'_b = \max r_{b,k}$, $k \in \mathcal{K}^{(\ell)}$, in order to ensure the URLLC's services allocation.

IV. NOMA FOR DOWNLINK SCHEDULING

In this section, we re-examine P0 under the assumption that it is possible to employ NOMA in the downlink to schedule different services, even at the mini-slot level [27]. We extend our analysis to the NOMA approach in order to discuss the potential gains that stem from the avoidance of conflicts, due to the superposition of services into the same resource block. NOMA has in the past been proposed as a competitive scheme to enhance throughput per resource block [26]; here we further motivate for it's employment as the means to mitigate conflicts in the allocation of resource blocks by allowing superposition of users, puncturing and preemptive scheduling can be avoided.

First, we consider the multiple NOMA (mNOMA) scenario where multiple users may share the same resource block. Therefore, P0 is reduced to a linear programming (LP), since this scheme allows overlapping amongst blocks, either fully or partially (on some of the minislots of the resource block). We refer to the corresponding optimization problem as P1, noting that the optimization variable is now a real number $x_{b,k} \in [0, 1]$ indicating the percentage of block $b \in \mathcal{B}$ assigned to the service $k \in \mathcal{K}$,

$$[\text{P1}] \quad \max_{x_{b,k} \in [0,1]} \sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{K}^{(c)}} r_{b,k} x_{b,k}, \tag{16}$$

$$\text{s.t.} \sum_{b \in \mathcal{B}} r_{b,k} x_{b,k} \geq q_k, \quad k \in \mathcal{K}^{(\ell)}, \tag{17}$$

$$\sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{K}} a_{b,i} x_{b,k} \leq \tilde{r}, \quad i \in \mathcal{I}. \tag{18}$$

$\tilde{r} \geq 1$ denotes the normalized sum throughput per block achieved with NOMA [32]. Note that the corresponding OMA constraint (6) is upper bounded to unity, pointing out a further gain in using NOMA due to the increase in per resource block utilization. However, as in this work we aim primarily at demonstrating the gains brought about due to conflict avoidance, in the numerical results presented in Section V we simply use $\tilde{r} = 1$.

A known issue of mNOMA is that error cascades in decoding can compromise performance; to alleviate such effects NOMA with user pairing has recently gained a lot of attention. In this framework, we implement the 2-user NOMA (2u-NOMA) scheme, since this approach provides lower decoding complexity, shorter delay and higher reliability in comparison to mNOMA [33]. Considering the power allocation problem in a downlink NOMA 2-user system, it has been proven in [33] that the achievable rate of the 2u-NOMA user with lower channel gain is equal to that of the OMA user when the power allocation is optimal and the remaining power is allocated to the strong 2u-NOMA user.

In contrast to the scheduling optimization problem as formulated in P1, 2u-NOMA allows overlapping amongst at most two blocks, either full or partial (of some mini-slots). In light of this, P1 is reformulated to a mixed integer programming (MIP) problem, by adding the supplementary binary variable $y_{b,k} \in \{0, 1\}$ to indicate weather block b is assigned to service k. To construct $y_{b,k}$, we let $\lceil * \rceil$. be the operator that maps a real number x to the smallest integer greater than or equal to x. The new MIP problem, referred to as [$\tilde{\text{P1}}$] follows,

$$[\tilde{\text{P1}}] \quad \max_{\{x_{b,k}, y_{b,k}\}} \sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{K}^{(c)}} r_{b,k} x_{b,k}, \tag{19}$$

$$\text{s.t. } y_{b,k} = \lceil * \rceil x_{b,k}, \quad k \in \mathcal{K}, \ b \in \mathcal{B} \tag{20}$$

$$\sum_{b \in \mathcal{B}} r_{b,k} x_{b,k} \geq q_k, \quad k \in \mathcal{K}^{(\ell)}, \tag{21}$$

$$\sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{K}} a_{b,i} x_{b,k} \leq \tilde{r}, \quad i \in \mathcal{I}, \tag{22}$$

$$\sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{K}} a_{b,i} y_{b,k} \leq 2, \quad i \in \mathcal{I}, \tag{23}$$

$$\sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{K}^{(\ell)}} a_{b,i} y_{b,k} \leq 1, \quad i \in \mathcal{I}, \tag{24}$$

$$\alpha_{b,i} r_{b,k'} x_{b,k'} \times 1_{x_{b,k'<1}} < \alpha_{b,i} r_{b,k''} x_{b,k''},$$
$$i \in \mathcal{I}, \quad b \in \mathcal{B}, \ k' \in \mathcal{K}^{(\ell)}, \ k'' \in \mathcal{K}^{(c)}. \tag{25}$$

The additional constraint (23) ensures that at most two overlapping blocks are allowed per mini-slot. Furthermore (24) forbids the overlapping URLLC resource blocks in order to avoid induced overheads in decoding. Finally, (25) ensures that if URLLC is overlapping with an eMBB resource block, the throughput of the eMBB resource block is higher.

(20) and (21) guarantee that the URLLC is always the weak user and thus, in the downlink NOMA, has to be decoded first [33], so that no extra latencies are introduced for URRL users due to the NOMA decoding order.

V. NUMERICAL RESULTS

In this section, we present numerical results for both OMA and NOMA schemes, for different 5G URLLC configurations and numerologies; fixed, multiple-fixed and flexible numerology. This exercise allows us to highlight the importance of flexible numerology, while motivating NOMA as a conflict mitigation approach. Here, we mainly focus on the conflicts aspect, rather than on deployment, feasibility or coordination issues, which are important enough to deserve an independent study. We then move on to a comparative analysis of the proposed heuristic Algorithms 1 (conflict aware, CA) and 2 (bin packing based, BPB) for OMA, to provide proof-of-concept for the potential of the proposed conflict aware scheduling.

We use the simulation setup given in [1], implemented based on the control channel overhead model for supporting the flexible numerology defined in [34] and considering the effect of guard band (i.e., of the cyclic prefix) on the achievable data rate, as modeled in [35]. The computation of the throughput per block $r_{b,k}$ relies on the configuration of block b (see Table 2), with a total number of nine multipath channel profiles [36], calculating the throughput based on the model introduced in [37]; for URLLC users the throughput values incorporate the delay constraints so that non-zero throughput is available only in these block in which the delay constraint is met. The throughput model also considers intersymbol-interference (ISI) depending on CP, and approximates the inter-channel interference (ICI) between the neighboring subbands of different numerologies.

In detail, the simulation parameters are given in Table 3. We assume a time-frequency grid, where each slot relies on a 2 msec and 2 MHz domain (i.e., of dimensions 16×11). As a result, this produces, for each slot, a set of $\mathcal{I} = \{1, \ldots, 176\}$ mini-slots and a corresponding set of $\mathcal{B} = \{1, \ldots, 549\}$ candidate blocks with respect to the numerology, where every candidate block consists of 4 elements of $\mathcal{I}$. The resource block details are given in Table 2. Blocks of shape 1 (corresponding to a shape of 4×1 minislots), $\mathcal{B}_1 \subset \mathcal{B}$, include a multitude of $|\mathcal{B}_1| = 143$ resource blocks. Blocks of shape 2 (2×2 minislots), $\mathcal{B}_2 \subset \mathcal{B}$, include a multitude of $|\mathcal{B}_1| = 150$ resource blocks. Blocks of shape 3 and 4 (1×4 minislots), $\mathcal{B}_3, \mathcal{B}_4 \subset \mathcal{B}$ include the same multitude of blocks $|\mathcal{B}_3| = |\mathcal{B}_4| = 128$. Moreover, the chosen latency tolerance and bit rate demands for the URRLC users are $\tau = \{0.5, 1, 1.5, 2\}$ msec and $q = \{16, 32, 64, 128, 256, 512\}$ kbits/sec (kbps), respectively. The latency tolerance for the eMBB users is fixed and equal to $\tau = 2$ msec. The SNR range is generated by numbers uniformly distributed in the interval [5, 30] dB. In the following, we refer to the "optimal shceduling" to denote the solutions provided by the Gurobi optimization solver, used as a benchmark for the evaluation of the optimality gap of the proposed heuristics.

TABLE 3. Simulation parameters.

Parameter	Value
Time frequency slot	2 ms and 2 MHz
$\mathcal{I}$	$\{1, \ldots, 176\}$
$\mathcal{B}$	$\{1, \ldots, 549\}$
τ	$\tau_k = \{0.5, 1, 1.5, 2\}$ ms, $k \in \mathcal{K}^{(\ell)}$ $\tau_k = 2$ ms, $k \in \mathcal{K}^{(c)}$
q	$\{16, 32, 64, 128, 256, 512\}$ kbps
SNR range	$U(5, 30)$ dB
block shapes	block shape 1 (4×1 minislots) block shape 2 (2×2 minislots) block shape 3 (1×4 minislots)
Traffic model	i. $\lvert\mathcal{K}^{(c)}\rvert = \lvert\mathcal{K}^{(\ell)}\rvert = 5$ ii. $\lvert\mathcal{K}^{(\ell)}\rvert \sim \text{Pois}(5)$, $\lvert\mathcal{K}\rvert = 10$ iii. $\lvert\mathcal{K}^{(\ell)}\rvert \sim \text{Pois}(5)$, $\lvert\mathcal{K}^{(c)}\rvert = 5$

Finally, we consider three scenarios for the arrival of the URLLC and the eMBB services on each slot: i) 5 URLLC and 5 eMBB constant users per slot, i.e., $|\mathcal{K}^{(c)}| = |\mathcal{K}^{(\ell)}| = 5$, ii) 10 users in total per slot, where the arrival rate of the URLLC services is a random variable following the Poisson distribution, i.e., $|\mathcal{K}^{(\ell)}| \sim \text{Pois}(5)$, $|\mathcal{K}| = 10$ and $|\mathcal{K}^{(c)}| = |\mathcal{K}| - |\mathcal{K}^{(\ell)}|$, and, iii) 5 constant eMBB users, $|\mathcal{K}^{(c)}| = 5$ and $|\mathcal{K}^{(\ell)}| \sim \text{Pois}(5)$. The outputs of all the simulation results are assessed over $M = 1000$ Monte Carlo iterations.

A. PERFORMANCE COMPARISON BETWEEN NOMA AND OMA SCHEDULING UNDER DIFFERENT NUMEROLOGIES

First, we compare the performance of OMA and NOMA schemes for different numerologies. In the case of fixed numerologies, shape 1 (horizontal), shape 2 (square) and shape 3 (vertical) type of blocks are considered separately. Furthermore, capturing a common scenario in practical systems [5], we define as the multiple-fixed numerology the one in which eMBB uses resource blocks of shape 1 (horizontal) and URLLC of shape 3 (vertical). Finally, in the case of flexible numerology all type of shapes, given in Table 2, are available to all services. In this subsection, we consider for the users arrivals, $|\mathcal{K}^{(\ell)}| = |\mathcal{K}^{(c)}| = 5$.

In Fig. 2, the sum bit rate for the eMBB services, $\mathcal{K}^{(c)}$ when applying the optimal i) OMA, ii) 2u-NOMA, and iii) mNOMA scheduling are shown. The NOMA sum bit rate gains with respect to the OMA are depicted with the lighter color in each bar. The latency tolerance and bit rate demands considered are $\tau = 1$ msec and $q = \{16, 32, 64, 128, 256, 512\}$ kbps, respectively, for five $\mathcal{K}^{(\ell)}$ and five $\mathcal{K}^{(c)}$ users. In all cases, as expected, flexible numerology significantly outperforms the fixed and multiple-fixed numerology. Moreover, multiple-fixed overpasses the per-

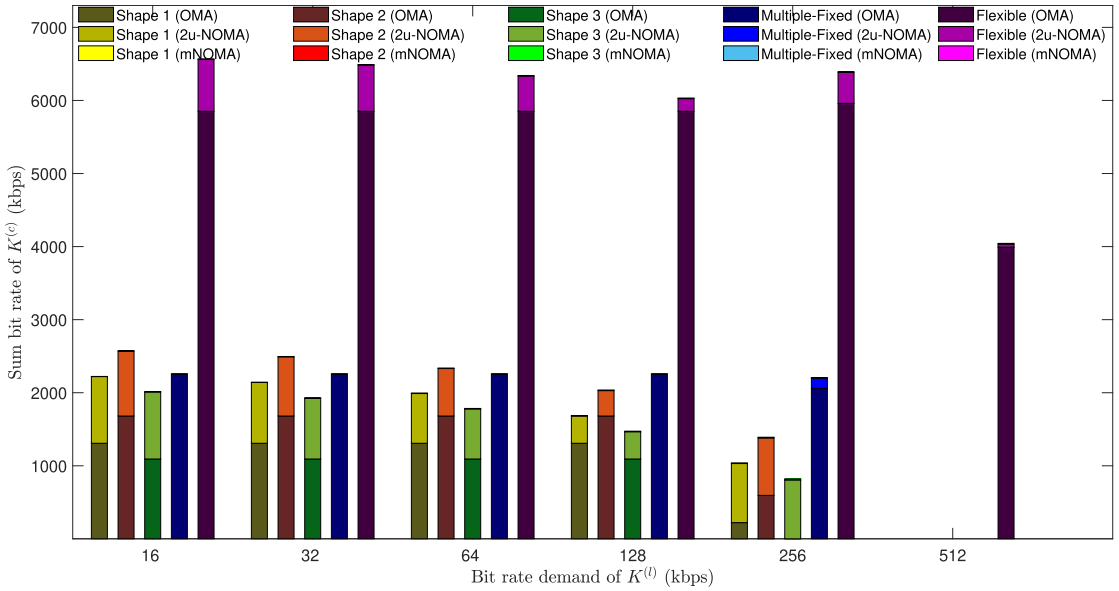


FIGURE 2. Sum bit rate of $\mathcal{K}^{(c)}$ services when employing OMA, 2-user NOMA (2u-NOMA) and multiple-user NOMA (mNOMA), considering fixed, multiple and flexible numerology, under several q_k data demands and delay tolerance of $k \in \mathcal{K}^{(l)}$ $\tau_k = 1$ msec. The lighter colors depict the NOMA sum bit rate gains in comparison to the OMA. Fixed and multiple-fixed numerologies result in infeasible outputs for $q_k = 512$ kbps.

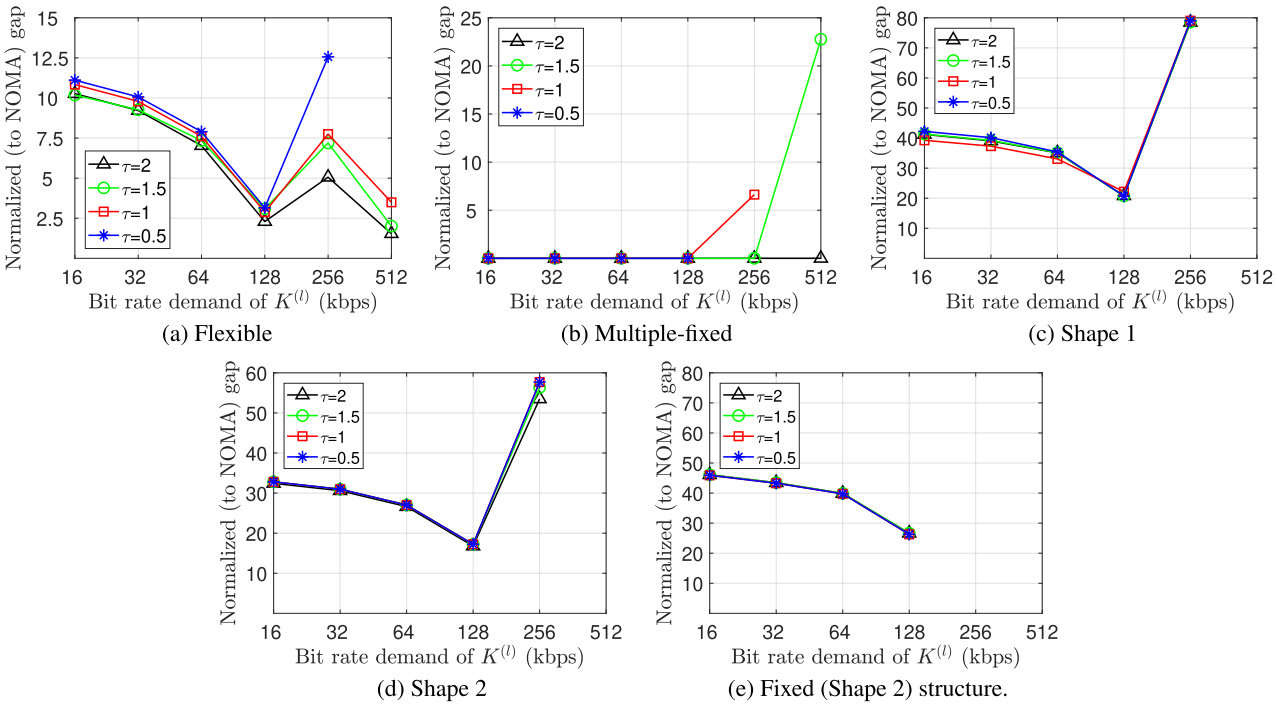


FIGURE 3. Normalized (to NOMA) gap of the sum bit rate of the $\mathcal{K}^{(c)}$ services between NOMA and OMA schemes. The y-axes measure percentages. Non existing values indicate infeasible solutions.

formance of fixed numerology in the OMA case. From these results it becomes apparent that flexible numerology in combination with NOMA can offer distinct gains across varying URLLC demands. Notably, as the URLLC demands increase, flexible numerology is the only approach that avoids infeasibility issues, i.e., not servicing all of URLLC demands.

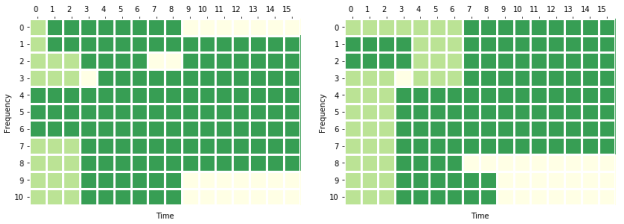
Focusing on the comparison between OMA and NOMA, both NOMA schemes consistently outperform OMA. More precisely, NOMA based scheduling is shown to increase

particularly the sum throughput of eMBB users under fixed numerology, although NOMA also improves the overall performance when using flexible numerology as well. On the other hand, NOMA does not affect the performance under multiple-fixed numerology; this is due to the fact that in the specific grid used in the simulations, overlapping of blocks is limited in the case of multi-fixed numerology. Furthermore, the gains in using NOMA are more accentuated in lower URLLC demands. Finally, the gains of mNOMA are negligible compared to these of 2u-NOMA for the specific grid; especially for lower q_k demands.

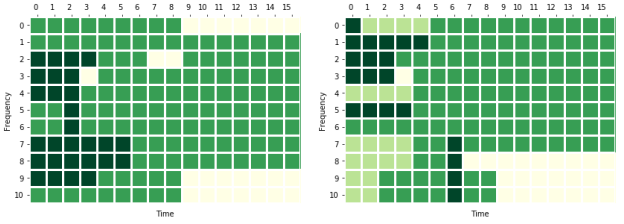
Furthermore, in Fig. 3, the normalized to NOMA performance gap between OMA and NOMA (expressed as a percentage) is shown, for different numerologies. The superiority of NOMA is reconfirmed both for fixed and flexible numerology, for different values of the URLLC latency tolerance $\tau_k = \{0.5, 1, 1.5, 2\}$ msec, $k \in \mathcal{K}^{(\ell)}$. Finally, in the case of flexible numerology, the lower the delay tolerance τ_k, the higher the gains in using NOMA as opposed to OMA. The performance fluctuations, illustrated in Fig. 3, are strongly related to the different values of the bit rate demands q_k, $k \in \mathcal{K}^{(\ell)}$. More precisely, after a close inspection of the simulation outputs, we came to the conclusion that the gap between the demand of a service $k \in \mathcal{K}^{(\ell)}$ and the achievable throughput of the block, in which the service is allocated, plays an important role. A higher gap between the two corresponds to a decisive reduction of the overall available throughput for the scheduling of the $\mathcal{K}^{(c)}$ services in the OMA case, which in turn offers a crucial advantage to the NOMA scheme that allows sharing of resource blocks.

In Fig. 4 examples of scheduling on the time-frequency grid is depicted in the case of OMA, 2u-NOMA and mNOMA, for $\tau_k = \{0.5, 1\}$ msec and $q_k = \{32, 256\}$ kbps, respectively, for all $k \in \mathcal{K}^{(\ell)}$. In the case of OMA and $q_k = 32$ depicted in Fig. 4(a), sharing of resource blocks is not allowed, while, in the case of NOMA, depicted in Figs. 4(c), (e), the opportunity of sharing resource blocks increases the sum throughput for eMBB, i.e., all of the blocks assigned to URLLC services are shared with these assigned to the eMBB services. Notice also that total number of blocks assigned to the eMBB services, for both mNOMA and 2u-NOMA, are the same; even if mNOMA concludes to a higher number of sharing blocks.

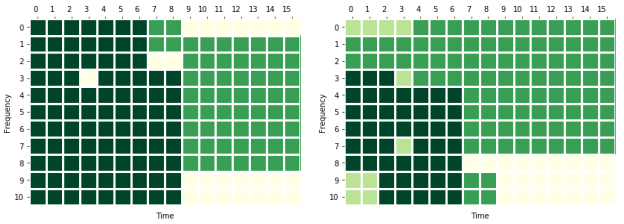
Similar outcomes are depicted in Fig. 4(b), when using OMA, and Figs. 4(d), (f), when using NOMA, in which a higher value of $q_k = 256$ is considered. In this case, though, not all of the blocks are assigned to eMBB services (for both NOMA scheduling) due to the higher URLLC demand.



(a) OMA scheduling, $\tau_k = 0.5$ ms and $q_k = 32$ kbps. (b) OMA scheduling, $\tau_k = 1$ ms and $q_k = 256$ kbps.



(c) 2u-NOMA scheduling, $\tau_k = 0.5$ ms and $q_k = 32$ kbps. (d) 2u-NOMA scheduling, $\tau_k = 1$ ms and $q_k = 256$ kbps.



(e) m-NOMA scheduling, $\tau_k = 0.5$ ms and $q_k = 32$ kbps. (f) m-NOMA scheduling, $\tau_k = 1$ ms and $q_k = 256$ kbps.

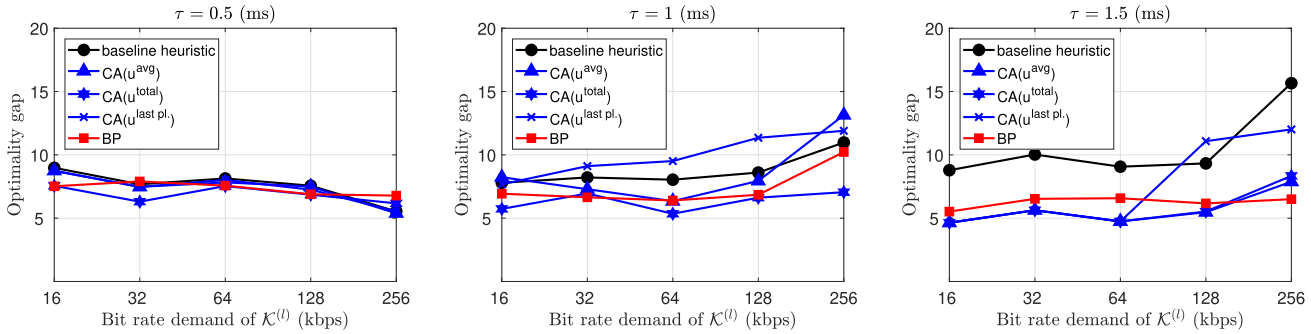
FIGURE 4. Resource allocation of URLLC (light green) and eMBB (green) services, for OMA (first row), 2u-NOMA (second row) and mNOMA (third row). Light yellow denotes zero throughput mini-slots. Dark green denotes sharing of mini-slots thanks to using NOMA.

B. PERFORMANCE OF PROPOSED HEURISTIC ALGORITHMS

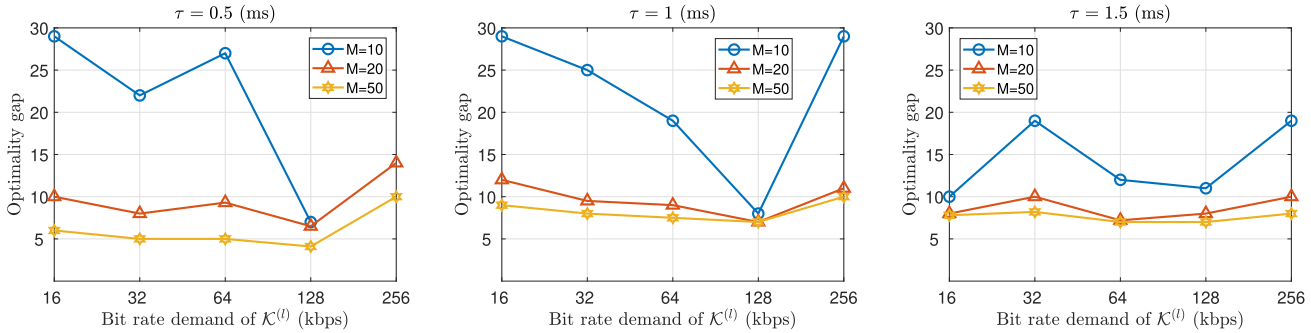
Although NOMA clearly outperforms OMA, its use might be prohibited by a number of factors, including the need for multiple decoding steps and the impact of imperfect successive interference cancelation (SIC). As a result, the evaluation of OMA scheduling approaches is paramount. In this subsection, we discuss the proposed heuristics. As a validation step, we first evaluate and compare the optimality gaps of the baseline heuristic (presented in [1]) and the proposed conflict aware heuristics with utilities ($\mathbf{u}^{total}$, $\mathbf{u}^{avg}$ and $\mathbf{u}^{last\ pl.}$), denoted by $CA(\cdot)$ with input one of the corresponding utility matrices, against the global optimum of P0. Then, we provide additional results with all proposed heuristics employing flexible numerology. For the above experimental results we assume that the URLLC users follow $|\mathcal{K}^{(\ell)}| \sim$ Pois(5) and the total amount of users in each slot is constant, $|\mathcal{K}| = 10$.

Fig. 5 depicts the optimality gap: i) of the baseline, the variations of the conflict-aware and the bin packing based approaches (first row), and, ii) of the LP-LD relaxation of P0 (second row), for several values of maximum sub-gradient iterations, with respect to the bit rate demand and the latency tolerance of the $\mathcal{K}^{(l)}$ services.

In the first row of Fig. 5 the conflict-aware and bin packing based heuristics are shown, in most cases, to outperform the baseline heuristic approach, especially for higher latency tolerance values, see Fig. 5(b) and (c), and to provide similar results for lower latency tolerance values, Fig. 5(a). The only exception is $CA(\mathbf{u}^{lastpl.})$ which outperforms the baseline approach only for $\tau = 1.5$ msec. More precisely, $CA(\mathbf{u}^{total})$,



(a) Baseline heuristic and CA heuristics.



(b) Baseline heuristic, using the LP-LD utility matrices, for several thresholds M.

FIGURE 5. a) Optimality gaps: a) of the baseline heuristic [1] and the variations of the conflict-aware heuristic CA, and, b) of the baseline LP-LD heuristic and thresholds for the sub-gradient iterations $M = \{10,\ 20,\ 50\}$. Against the global optimum of P0, for latency tolerance values $\tau_k = \{0.5,\ 1,\ 2\}$ ms. The y-label express the relative deviation to the optimum, expressed as percentage.

$CA(\mathbf{u}^{avg})$ BP clearly outperform all other approaches, maintaining an optimality gap below to 10% for $\tau_k = 0.5$ msec and close to 7% for $\tau_k = \{1, 1.5\}$ msec.

The second row of Fig. 5 depicts the optimality gap of the LP-LD heuristic solutions, for various threshold values $M = \{10, 20, 50\}$ of the maximum sub-gradient iterations, against the global optimum. Incorporating the utility matrices $\mathbf{u}_{LP},\ \mathbf{u}_{LD} \in \mathbb{R}_{\mathcal{B}\times\mathcal{K}}$ leads to similar results (not depicted for compactness). As it is expected, higher threshold values of M lead to a further reduction in the optimality gap, at the cost of a higher computational time. The choice of $M = 10$ results on very high optimality gaps, near to 20% in most cases. On the other hand for $M = 20$ and $M = 50$ the heuristics are shown to maintain the optimality gap close and lower than 10% for the chosen latency tolerance values. Note that the optimality gaps of the proposed heuristics are comparable to that of the LP-LD variations in all cases, as it can be seen by comparing the two rows of Fig. 5. On the other hand, the reduction of the optimality gap using LP-LD utility matrices comes with a significant increase of the computational time.

Furthermore, we utilize our implementation to quantify the performance of the optimal and the heuristic approaches, in terms of processing cost. The computational time is measured on a Lenovo IdeaPad 510-15IKB laptop, with an Intel Core i7-7500U @ 2.70 GHz processor and 12 GB RAM. In Fig. 6, we depict the processing cost of i) the optimal solution; ii) the baseline heuristic variations (without the usage of the LP-LD utilities); iii) the bin packing based approach; and,

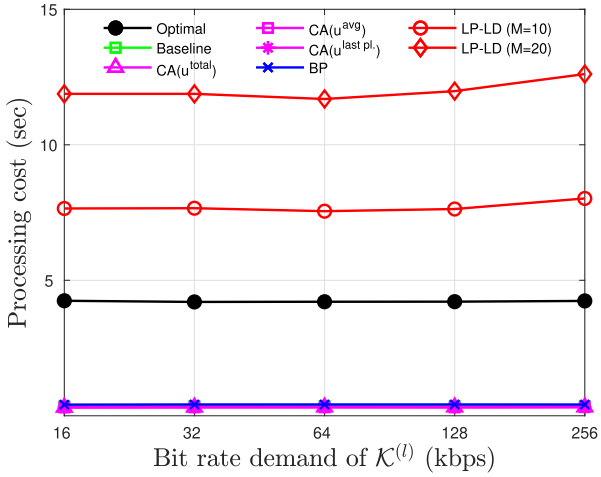


FIGURE 6. The processing cost of: i) the optimal, ii) the baseline heuristic variations, iii) the bin packing based approach, and, iv) the LP-LD ($M = 20$), for $\tau = 1$ msec and $q_k = \{16,\ 32,\ 64,\ 128,\ 258\}$ (kbps).

iv) the LP-LD heuristic with threshold value $M = \{10,\ 20\}$, for $q_k = \{16, 32, 64, 128, 256\}$ and a conventional latency tolerance value $\tau = 1$ ms. As shown, the LP-LD solution is much more computational intensive than other heuristic approaches, even compared to the optimal solution. Note that higher threshold values increase drastically the processing cost, e.g., for $M = 50$ the processing cost is of 22 sec. On the other hand, the processing cost of the bin packing

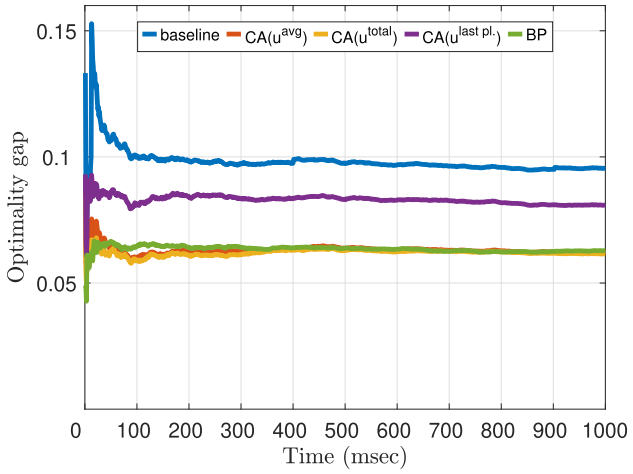


FIGURE 7. Cumulative optimality gap over time, when the latency tolerance and the bit rate demands of the $\mathcal{K}^{(\ell)}$ users are randomly chosen, for each time instance, from $q_k = \{16, 32, 64, 128, 256\}$ kbps and $\tau = \{0.5, 1, 1.5, 2\}$ msec.

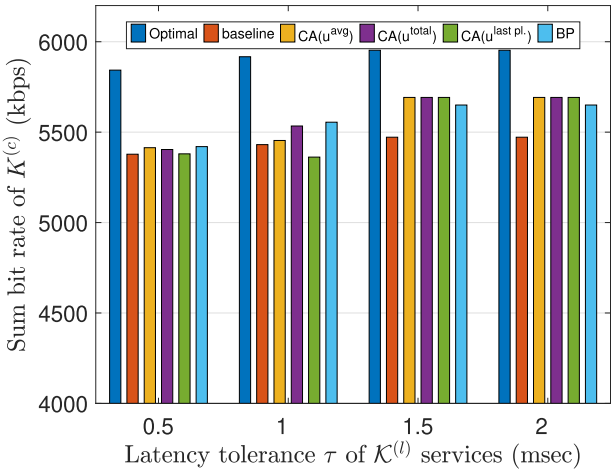


FIGURE 9. Sum bit rate of $\mathcal{K}^{(c)}$ services for various values for the latency tolerance of $\mathcal{K}^{(\ell)}$ services, when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 32 kbps.

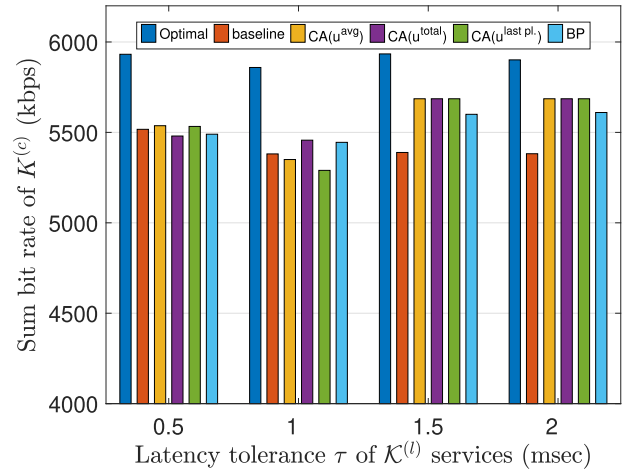


FIGURE 8. Sum bit rate of $\mathcal{K}^{(c)}$ services for various values for the latency tolerance of $\mathcal{K}^{(\ell)}$ services, when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 16 kbps.

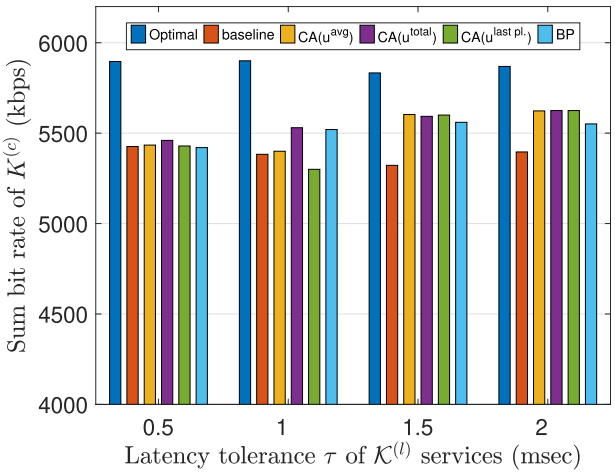


FIGURE 10. Sum bit rate of $\mathcal{K}^{(c)}$ services for various values for the latency tolerance of $\mathcal{K}^{(\ell)}$ services, when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 64 kbps.

and the conflict-aware heuristics is between 0.3 and 0.5 sec, noting that the complexity of the conflict-aware and bin packing based heuristics is of $\mathcal{O} = (N \log N)$. Based on the computational complexity comparison, we exclude from the next experiments the approaches including the calculation of the LP-LD utility matrices.

Finally, in Fig. 7 we illustrate the cumulative optimality gap over time, where the throughput demands and the latency tolerance of the URLLC services are randomly chosen (in each time instance) from the vectors $q_k = \{0.016, 0.032, 0.064, 0.128, 0.256\}$ kbps and $\tau_k = \{0.5, 1, 1.5, 2\}$ msec, respectively. The results demonstrate the superior performance of the conflict aware and the bin packing based heuristics, i.e., the optimality gap converges to 6% for BP, $CA(\mathbf{u}^{total})$ and $CA(\mathbf{u}^{avg})$, while, the optimality gap of the baseline approach converges to 10%.

In the next set of experiments, we assume that the total number of the URLLC services follow the Poisson distribution, $|\mathcal{K}^{(\ell)}| \sim \text{Pois}(5)$ and the eMBB services are constant per slot, $|\mathcal{K}^{(c)}| = 5$. Figs. 8 - 10 show that the performance of the heuristic algorithms compares well with the global optimum (obtained through Gurobi solvers), while keeping the complexity very low. Note that the proposed algorithms, exceed the performance of the baseline heuristic, especially for $\tau >$ 0.5 msec, verifying the results of the previous experiments Figs. 5(a), 7. This showcases that indeed, the reformulation of the optimal scheduling as a conflict minimization problem is highly pertinent and allows shedding light on how to jointly address the constraints (5) and (6) of P0. It is also noteworthy that more elaborate heuristics could be proposed in the same context, by looking at algorithms with lower optimality gaps to the optimal bin packing solution.

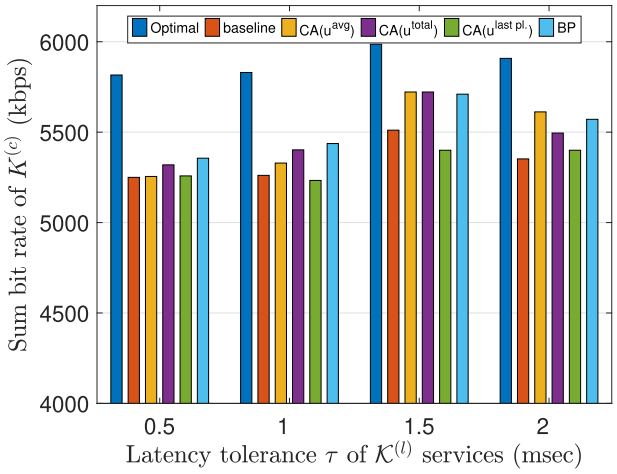


FIGURE 11. Sum bit rate of $\mathcal{K}^{(c)}$ services for various values for the latency tolerance of $\mathcal{K}^{(\ell)}$ services, when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 128 kbps.

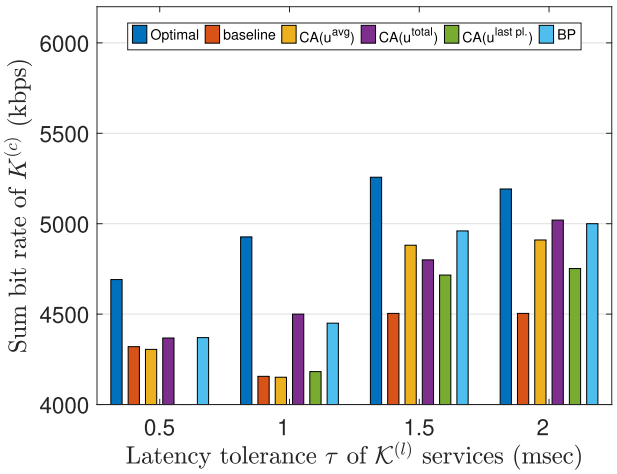


FIGURE 12. Sum bit rate of $\mathcal{K}^{(c)}$ services for various values for the latency tolerance of $\mathcal{K}^{(\ell)}$ services, when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 256 kbps.

The same conclusions can be reached in Figs. 11 and 12 for URLLC demands of 128 and 256 kbps, respectively. In these cases, all the conflict-aware choices exceed the performance of the baseline heuristic; the choice of $CA(\mathbf{u}^{lastpl\cdot})$ metric is the only one with lower performance to that of baseline heuristic for $\tau = 1$ ms and for $q_k = 256$ kbps and $\tau = 0.5$ msec. Also, the $CA(\mathbf{u}^{total})$, $CA(\mathbf{u}^{avg})$ and the *BP* heuristics provide similar performance rates.

Finally, in Fig. 13 we illustrate the cumulative optimality gap over time, considering random q_k and τ_k values for each $k \in \mathcal{K}^{(\ell)}$ (according to Fig. 7), further assuming that $|\mathcal{K}^{(\ell)}| \sim \text{Pois}(6)$ and $|\mathcal{K}^{(c)}| = 5$. The superior performance of the proposed heuristics is reconfirmed, especially for the *BP* and the $CA(\mathbf{u}^{avg})$, $CA(\mathbf{u}^{total})$ approaches, which significantly reduce the gap; converge to 6% when the baseline heuristic converges to 11%.

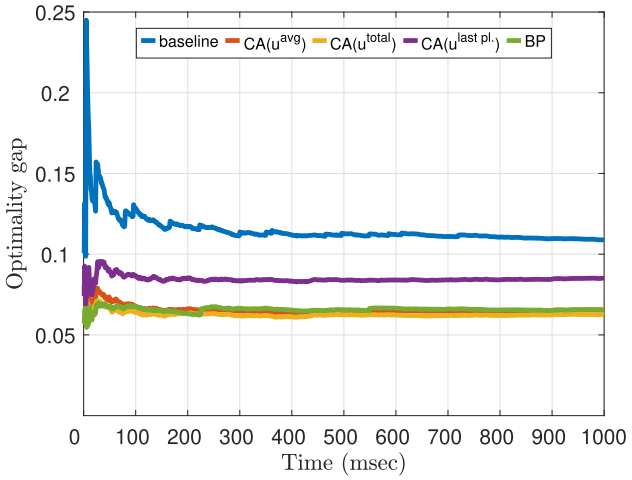


FIGURE 13. Cumulative optimality gap over time, when the total number of the URLLC users follow the Poisson distribution and the eMBB services are constant. The latency tolerance and the bit rate demands of the $\mathcal{K}^{(\ell)}$ users are randomly chosen, for each time instance, from $q_k = \{16, 32, 64, 128, 256\}$ kbps and $\tau = \{0.5, 1, 1.5, 2\}$ msec.

VI. CONCLUSION AND FUTURE WORK

In 5G and beyond networks, URLLC services will coexist with eMBB services, giving rise to challenging layer 2 scheduling. To address the latter, we have reformulated the standard eMBB throughput maximization problem as an equivalent conflict minimization, which points at minimizing the overall amount of conflicts. Building on this premise, two lightweight and efficient scheduling approaches were proposed; a family of conflict-aware heuristics that employ conflict aware utilities and a heuristic inspired by the bin packing problem. In addition to the proposed scheduling using orthogonal multiple access (OMA), we further proposed the use of non-orthogonal multiple access (NOMA) to mitigate conflicts. We investigated the potential advantages of allowing for non-orthogonal sharing of radio resources with flexible numerology and frame structure. The intuition for NOMA's superior performance, as a result of alleviating conflicts, was demonstrated to hold; importantly, NOMA can potentially offer significant advantages particularly in the case of ultra-low latency constraints for the URLLC users. Extensive simulations were performed for URLLC services with different QoS requirements both for OMA and NOMA scenarios. The simulation results showed that i) all of the proposed heuristics have near-optimal performance, demonstrating that conflict minimization is indeed key to layer 2 scheduling and that there are significant gains in terms of resource utilization, when employing NOMA.

In the future work we will extend the existing approach targeting on a wider range of performance metrics for the eMBB users, e.g., minimum expected achieved range and fairness. In this framework, we would also consider scheduling schemes for the uplink.

ACKNOWLEDGMENT

The authors would like to thank the authors of [1] for kindly sharing their simulation codes in IEEE DataPort [2].

REFERENCES

[1] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579–2582, Dec. 2018.
[2] L. You, Q. Liao, N. Pappas, and D. Yuan, "2D resource allocation," *IEEE Dataport*, Aug. 2018, doi: 10.21227/ch8e-x385.
[3] *Setting Scene for 5G: Opportunities challenges*, Int. Telecommun. Union (ITU), Geneva, Switzerland, 2018.
[4] *NR; Physical Channels and Modulation, Release 16*, document (TS) 38.211 V 16.1.0, 3GPP, 2020.
[5] Y. Sadi, S. Erkucuk, and E. Panayirci, "Flexible physical layer based resource allocation for machine type communications towards 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.
[6] Y. Abiko, T. Saito, D. Ikeda, K. Ohta, T. Mizuno, and H. Mineno, "Flexible resource block allocation to multiple slices for radio access network slicing using deep reinforcement learning," *IEEE Access*, vol. 8, pp. 68183–68198, 2020.
[7] F. Mason, G. Nencioni, and A. Zanella, "Using distributed reinforcement learning for resource orchestration in a network slicing scenario," 2021, *arXiv:2105.07946*.
[8] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.
[9] 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 Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.
[10] O. Semiari, W. Saad, M. Bennis, and M. Debbah, "Integrated millimeter wave and sub-6 GHz wireless networks: A roadmap for joint mobile broadband and ultra-reliable low-latency communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 109–115, Apr. 2019.
[11] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, Dec. 2018.
[12] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Netw.*, vol. 32, no. 2, pp. 24–31, Mar./Apr. 2018.
[13] M. Almekhlafi, M. Chraiti, M. A. Arfaoui, C. Assi, A. Ghrayeb, and A. Alloum, "A downlink puncturing scheme for simultaneous transmission of URLLC and eMBB traffic by exploiting data similarity," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13087–13100, Dec. 2021.
[14] Y. Huang, Y. T. Hou, and W. Lou, "A Deep-Learning-based link adaptation design for eMBB/URLLC multiplexing in 5G NR," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Vancouver, BC, Canada, May 2021, pp. 1–10.
[15] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, pp. 45674–45688, 2020.
[16] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.
[17] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.
[18] 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 Netw.*, vol. 32, no. 2, pp. 8–15, Mar./Apr. 2018.
[19] A. Pradhan and S. Das, "Joint preference metric for efficient resource allocation in co-existence of eMBB and URLLC," in *Proc. Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Bengaluru, India, Jan. 2020, pp. 897–899.
[20] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Toronto, ON, Canada, Sep. 2017, pp. 1–6.
[21] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "EMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.
[22] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.
[23] J. Li and X. Zhang, "Deep reinforcement learning-based joint scheduling of eMBB and URLLC in 5G networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1543–1546, Sep. 2020.
[24] N. Ferdosian, S. Skaperas, A. Chorti, and L. Mamatas, "Conflict-aware multi-numerology radio resource allocation for heterogeneous services," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.
[25] M. Mitev, A. Chorti, M. Reed, and L. Musavian, "Authenticated secret key generation in delay-constrained wireless systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, pp. 1–29, Dec. 2020.
[26] A. T. Abusabah and H. Arslan, "NOMA for multinumerology OFDM systems," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–9, Aug. 2018.
[27] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.
[28] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Netw.*, vol. 31, no. 4, pp. 8–14, Jul./Aug. 2017.
[29] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.
[30] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.
[31] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, vol. 21, 6th ed. Berlin, Germany: Springer-Verlag, 2018.
[32] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.
[33] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019.
[34] H. Miao and M. Faerber, "Physical downlink control channel for 5G new radio," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Oulu, Finland, Jun. 2017, pp. 1–5.
[35] A. Yazar and H. Arslan, "A flexibility metric and optimization methods for mixed numerologies in 5G and beyond," *IEEE Access*, vol. 6, pp. 3755–3764, 2018.
[36] *Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception, Release 14*, document (TS) 36.101 V 14.3.0, 3GPP, 2017.
[37] M. Batariere, K. Baum, and T. P. Krauss, "Cyclic prefix length analysis for 4G OFDM systems," in *Proc. IEEE 60th Veh. Technol. Conf. (VTC-Fall)*, Los Angeles, CA, USA, Sep. 2004, pp. 543–547.

SOTIRIS SKAPERAS (Member, IEEE) received the B.Sc. degree in mathematics and the M.Sc. degree in statistics and modeling from the Aristotle University of Thessaloniki, Greece, in 2013 and 2016, respectively, and the Ph.D. degree in resource management (5G networks) from the Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece, in 2021. He has been a Postdoctoral Researcher at the CY Cergy Paris University, France, since June 2021. His research interests include resource allocation/load balancing for 5G networks using time-series/change point analysis and wireless communication security for 5G networks.

NASIM FERDOSIAN (Member, IEEE) received the M.Tech. degree (Hons.) in computer science from Jawaharlal Nehru Technological University, India, in 2010, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, in 2018. She is currently a Research Associate with Curtin University, Australia. She has authored a number of articles in high impact factor scientific journals. Her current research interests include new radio resource management algorithms for the next generations of cellular networks and quality-of-service in wireless communications and cellular networks.

ARSENIA CHORTI (Senior Member, IEEE) received the M.Eng. degree in electrical and electronic engineering from the University of Patras, Greece, the D.E.A. degree in electronics from the University Pierre et Marie Curie, Paris VI, France, and the Ph.D. degree in electrical engineering from Imperial College London, U.K., in November 2005. She was a Postdoctoral Researcher at the University of Southampton, U.K., the Technical University of Crete, Greece, and the University College London, U.K., from 2005 to 2008. She was a Senior Lecturer in communications at Middlesex University, U.K., from December 2008 to April 2010. From 2010 to 2013, she was a Marie Curie IOF Researcher with Princeton University, Princeton, NJ, USA, and the Institute of Computer Science-FORTH, Greece. From 2013 to 2017, she was a Lecturer with the University of Essex, U.K. She is currently a Professor with the Ecole Nationale Superieure de l'Electronique et des ses Applications (ENSEA); the Head of the Information, Communications and Imaging (ICI) Group, ETIS Laboratory UMR 8051; and a Visiting Research Fellow with Princeton University and the University of Essex. Her research interests include wireless communications and wireless system security for 5G and 6G, with a particular focus on physical layer security, context aware security, multi-factor authentication protocols, 5G/6G and the IoT, anomaly detection, machine learning for communications, nonorthogonal multiple access (NOMA), and faster than Nyquist signaling. She is a member of IEEE INGR on Security and the IEEE P1951.1 Standardization Workgroup (Smart Cities).

LEFTERIS MAMATAS (Member, IEEE) received the Diploma and Ph.D. degrees from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2003 and 2008, respectively. He was a Researcher with the University College London, London, U.K.; the Space Internetworking Center/Democritus University of Thrace, Greece; and the DoCoMo Euro-Labs, Munich, Germany. He is currently an Assistant Professor with the Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece, where he leads the Softwarized and Wireless Networks Research Group. He has published more than 60 articles in international journals and conferences. He participated in many international research projects, such as NECOS (H2020), FED4FIRE+OC4 (H2020), WiSHFUL OC2 (H2020), MONROE OC2 (H2020), Dolfin(FP7), UniverSELF (FP7), and Extending Internet into Space (ESA). His research interests include software-defined networks, the Internet of Things, 5G networks, and multi-access edge computing. He served as the General Chair for the WWIC2016 Conference and the INFO-COM SWFAN 2016 Workshop and the TPC Chair for the INFOCOM SWFAN 2017, E-DTN 2009, and IFIP WWIC 2012 conferences/workshops. He is a Guest Editor of the *Ad Hoc Networks* journal (Elsevier).