2021 IEEE International Conference on Communications Workshops (ICC Workshops) | 978-1-7281-9441-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCWorkshops50388.2021.9473664

Conflict-Aware Multi-Numerology Radio Resource Allocation for Heterogeneous Services

Nasim Ferdosian
ETIS UMR 8051
CY-Tech, ENSEA, CNRS
Cergy 95000 France
nasim.ferdosian@ensea.fr

Sotiris Skaperas
Dep. of Applied Informatics
University of Macedonia
Thessaloniki, Greece
sotskap@uom.edu.gr

Arsenia Chorti
ETIS UMR 8051
CY-Tech, ENSEA, CNRS
Cergy 95000 France
arsenia.chorti@ensea.fr

Lefteris Mamatas
Dep. of Applied Informatics
University of Macedonia
Thessaloniki, Greece
emamatas@uom.edu.gr

***Abstract*—5G new radio (NR) introduced flexible numerology to accommodate applications with varying quality of service (QoS) requirements. However, optimizing the scheduling of services with varying delay and throughput QoS constraints remains a challenging task. Under existing proposals, supporting ultra-reliable low-latency communication (URLLC) services comes at the cost of reduced throughput offered to enhanced mobile broadband (eMBB) users. In this paper, we propose an efficient, low complexity scheduling of radio resources for URLLC when coexisting with eMBB services. We re-formulate the standard eMBB throughput maximization problem as an equivalent conflict minimization with URLLC and prove that this equivalent objective can be treated as a Bin Packing optimization problem. Moreover, in order to further increase the efficiency of resource utilization, non-orthogonal multiple access (NOMA) is also investigated for URLLC and eMBB coexistence. The superior performance of NOMA, with the superposition of services over the same resource blocks, is due to alleviating conflicts, as shown by an extensive set of numerical results.**

***Index Terms*—Flexible numerology, URLLC traffic.**

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 varying quality of service (QoS) requirements and characteristics. The 3rd generation partnership project (3GPP) standardized 5G in the form of a novel radio interface technology, referred to as new radio (NR) [1]. 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 [2], [3], [4]. The major challenges related to radio resource optimization for ultra-reliable low-latency communication (URLLC) systems are described in [5].

In 5G and beyond, URLLC services with extreme delay constraints will coexist with enhanced mobile broadband (eMBB), that require very high bit rates (Gigabits per second) and have moderate latency requirements. On the other hand, URLLC services are expected to have lower traffic volumes than eMBB services. The design of radio resource allocation strategies for URLLC traffic when coexisting with eMBB has been a focal point of recent research efforts [6]. In [7], [8] resource allocation strategies for the coexistence of URLLC and eMBB were proposed based on a "puncturing" framework: according to this, eMBB traffic was scheduled initially at the beginning of the slots; upon arrival of URLLC traffic, the latter was prioritized and dynamically overlapped at mini-slots of ongoing eMBB transmissions (which were punctured). These approaches have been shown to result in significant losses in terms of data rates for eMBB services [9].

Alternatively, the authors in [10] studied the resource allocation of eMBB and URLLC services, without using puncturing mechanisms to schedule resources. A flexible numerology and frame structure was considered by defining a time-frequency resource grid, containing four different 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.

In this work, we first propose a conflict-aware, multi numerology radio resource allocation algorithm to maximize scheduling efficiency for URLLC when coexisting with eMBB services. The proposed scheduling approach results from reformulating the standard eMBB throughput maximization problem in an equivalent form in which the objective is to minimize conflicts with URLLC in terms of resource allocation. This new problem is shown to be solved by jointly minimizing the placements of URLLC services in the time-frequency resource grid and the aggregate conflict, which can be treated as a specific instance of bin packing optimization. Simulation results show that a heuristic scheduling algorithm of near-linear complexity, provides a quick, lightweight and efficient solution to resource allocation scheduling in URLLC and eMBB coexistence.

Moreover, having shed light to the importance of minimizing conflicts between different services, the utilization of non-orthogonal multiple access (NOMA) schemes naturally emerges as a competitive candidate [11], [12]. NOMA allows for the superposition of services, even at the mini-slot level by employing superposition coding at the transmitter and successive interference cancellation at the receivers [13], [14]. NOMA has in the past been proposed as a competitive scheme

978-1-7281-9441-7/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: University of Macedonia. Downloaded on July 30,2021 at 11:55:01 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Resource Blocks in Flexible Numerology

	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

to enhance throughput per resource block [15]; in the present work we provide further motivation for it's employment in beyond 5G networks (B5G) as the means to mitigate conflicts in the allocation of resource blocks, i.e., in layer 2 scheduling. We provide an extensive set of numerical results that show the significant gains in terms of eMBB throughput when adopting NOMA in a flexible numerology setting.

The paper is organized as follows. The resource allocation optimization problem is described in Section II, along with the equivalent formulation as a conflict minimization. A near-linear complexity heuristic algorithm is proposed in Section III, inspired by a greedy heuristic solution to the bin packing problem, along with the problem re-formulation when using NOMA. Section IV presents numerical results showing the performance of the heuristic as well as the superiority of NOMA for URRLC and eMBB coexistence. Finally, conclusions are drawn in Section V.

II. Problem Formulation

Following the system model in [10], we consider a base station serves both throughput hungry users (eMMB) and ultra-low latency users (URLLC). The objective is to find the resource allocation in the time-frequency grid that maximizes the sum throughput of the former, while satisfying the throughput demands and latency constraints of the latter.

$\mathcal{K}$ denotes the set of all services, $\mathcal{K}^{(c)}$ the set of eMBB users, $\mathcal{K}^{(\ell)}$ the set of URLLC users, q_k and τ_k are respectively the throughput demand and maximum tolerant latency of service $k \in \mathcal{K}^{(\ell)}$. $\mathcal{B}$ is the set of all possible resource blocks according to the numerology employed and finally, $\mathcal{I}$ denotes the set of all mini-slots. We utilize the 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 $\alpha_{b,i} = 0$. Furthermore, we denote by $r_{b,k}, b \in \mathcal{B}, k \in \mathcal{K}$ the throughput of each resource block, under the constraint that the latency constraint is met, i.e.,

$$r_{b,k} = \{\text{Capacity of block } b \text{ for service } k\} \times \mathbf{1}_{\{\tau_k - t_b > 0\}} \tag{1}$$

where t_b is the end time of block b and $\mathbf{1}_{\{x\}}$ is the indicator function for the logical proposition x. Finally, 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 $x_{b,k} = 0$. Table I represents four most widely accepted block shapes on the 5G NR. According to the flexible numerology, $\mathcal{K}^{(c)}$ (eMBB) and $\mathcal{K}^{(\ell)}$ (URLLC) services have no restrictions and they are able to choose any of the given shapes.

The standard scheduling optimization problem is to maximize 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 blocks. The formal problem formulation is given as,

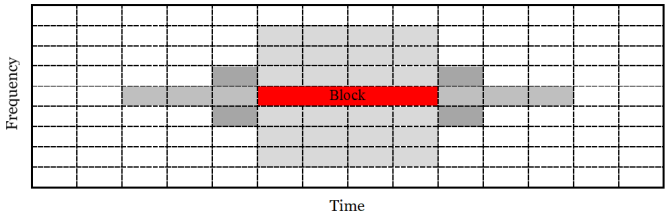

Fig. 1: Resource allocation of a candidate block and its corresponding conflicts; vertical blocks (light grey), horizontal blocks (grey) and square blocks (dark grey).

$$\text{[P0]} \quad \max_{x_{b,k}\in\{0,1\}} \sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}^{(c)}} r_{b,k}x_{b,k}, \tag{2}$$

$$\text{s.t.} \quad \sum_{b\in\mathcal{B}} r_{b,k}x_{b,k} \geq q_k, \quad k \in \mathcal{K}^{(\ell)}, \tag{3}$$

$$\sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}} \alpha_{b,i}x_{b,k} \leq 1, \quad i \in \mathcal{I}. \tag{4}$$

In [10], it is proved that the combinatorial problem P0 is an NP-hard partition problem. A heuristic algorithm, named LP-LD, was proposed based on the linear programming (LP) and the Lagrange dual (LD) relaxation of P0, without considering the *impact* of the $\mathcal{K}^{(\ell)}$ services allocation to the consequent allocation of the $\mathcal{K}^{(c)}$ services. The complexity of the LP-LD algorithm was shown to be $\mathcal{O}(|\mathcal{B}||\mathcal{K}|\log(|\mathcal{B}||\mathcal{K}|))$ ignoring the high complexity of the computation of utility matrices; demand the usage of optimization solvers, while the dual LP-LD approach also applies a sub-gradient method.

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. We consider the number of generated "conflict" by any specific URLLC or eMBB resource block placement. To illustrate the idea, Fig. 1 depicts all the "conflicts" that arise from an arbitrary block placement, shown in red; the specific block allocation (in red) *forbids* any other block allocation in the sketched neighborhood (in grey). In light of this, even if a particular resource block might have maximum throughput, its allocation could be suboptimal due to the losses caused by the generated forbidden placements around it. To evaluate the impact of (4), we define any conflict (overlapping) of resource blocks as,

$$c_{b,p} = \begin{cases} 1, & \text{if } \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{B}} (\alpha_{b,i} + \alpha_{p,i}) > 1,\ i \in \mathcal{I},\ b \neq p \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

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

$$\sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}^{(c)}} r_{b,k}x_{b,k} = R_{total} - \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{B}} \sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k}, \tag{6}$$

where R_{total} denotes the maximum sum throughput of the whole resource grid with respect to $\mathcal{K}^{(c)}$ and the second

Authorized licensed use limited to: University of Macedonia. Downloaded on July 30,2021 at 11:55:01 UTC from IEEE Xplore. Restrictions apply.

triple sum represents the losses in $\mathcal{K}^{(c)}$ throughput because of the conflicts generated by the placements of all services. As a result, the maximization of (2) is equivalent to the minimization of the aggregate conflict, i.e.,

$$\max_{x_{b,k}\in\{0,1\}} \left(R_{total} - \sum_{b\in\mathcal{B}}\sum_{p\in\mathcal{B}}\sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k} \right) \Leftrightarrow$$
$$\min_{x_{b,k}\in\ \{0,1\}} \sum_{b\in\mathcal{B}}\sum_{p\in\mathcal{B}}\sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k}. \quad (7)$$

Since the R_{total} is constant (for any given grid and throughput values), the maximization problem may be reduced to the minimization of the potential conflicts. We also note that:

$$\mathbb{E}\left[\sum_{b\in\mathcal{B}}\sum_{p\in\mathcal{B}}\sum_{k\in\mathcal{K}} c_{b,p}x_{p,k}r_{b,k}\right] = |\mathcal{C}|\,\bar{r}, \quad (8)$$

where $\mathbb{E}[\cdot]$ denotes expectation, $\mathcal{C}$ is the set of conflicts when all resource blocks have the same throughput $\bar{r} = \mathbb{E}[r_{b,k}]$ and $|\cdot|$ denotes cardinality; i.e., in the long term we need *on average* to minimize the *number* of conflicts. The minimization of the number of conflicts can be taken through the minimization of the number of URLLC placements, which points to a formulation of the scheduling problem as a bin packing optimization problem [16]. In the following Section, inspired by a near-linear complexity greedy heuristic for the bin packing, we propose a novel lightweight scheduling approach that is shown through numerical results to be also very resource efficient.

III. Heuristic Algorithm and Conflict Resolution by Using NOMA

A. Heuristic Inspired from Bin Packing Optimization

The proposed scheduling heuristic that accounts for conflicts is summarized in Algorithm 1, jointly minimizing the number of $\mathcal{K}^{(\ell)}$ resource allocations (placements) and throughput losses for $\mathcal{K}^{(c)}$ users. Allocation of resources to $\mathcal{K}^{(\ell)}$ services and $\mathcal{K}^{(c)}$ services is treated sequentially, with the former being served first to meet the latency requirements. In the following, the vector $\mathbf{e}$ of length $|\mathcal{B}|$ has as elements the aggregated throughput losses for each allocation of a block $b \in \mathcal{B}$, i.e.,

$$e_b = \sum_{p\in\mathcal{B}}\sum_{k\in\mathcal{K}^{(c)}} c_{b,p}r_{b,k}. \quad (9)$$

For each $k \in \mathcal{K}^{(\ell)}$ we generate M categories (bins) with decreasing fractional sizes with respect to $q_k, k \in \mathcal{K}^{(\ell)}$, i.e., category $i \in \{1,\dots,M\}$ is defined as the set of all resource blocks $b \in \mathcal{B}$ for which the ceiling of the ratio of the service demand 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.

$$Cat^iU^k = \left\{ b : \left\lceil \frac{q_k}{r_{b,k}} \right\rceil = i, \forall b \in \mathcal{B} \setminus \{Cat^jU^k\}_{j=1,\dots,i-1} \right\},$$
$$k \in \mathcal{K}^{(\ell)}, i \in \{1,\dots,M\}, \quad (10)$$

Algorithm 1 Bin Packing Resource Allocation Algorithm

Input: throughput matrix $\mathbf{r} = [r_{b,k}]$, $b \in \mathcal{B}, k \in \mathcal{K}$, aggregated-throughput-loss vector $\mathbf{e}$, demand vector of URLLC services $\mathbf{q}$, set of all available resource blocks $\mathcal{B}$.
Output: Block-service assignment $\mathbf{s}$.

```
for k = 1 to |q| do
    create the following categories:
    for i = 1 to M do
        Cat^i U^k = all resource blocks b ∈ B where
        ⌈q_k / r_{b,k}⌉ = i;
        Check pairwise conflicts among categorized blocks
        and remove the blocks with the higher aggregated-
        throughput-loss;
    end for
end for
Phase (K^(ℓ) resource allocation):
for i = 1 to M do
    select the Cat^i U^k which has the least number of blocks;
    if (|Cat^i U^k| ≥ i and q_k is not already met) then
        B' ← (select i number of blocks in Cat^i U^k with the
        least aggregated-loss-value);
        s ← s ∪ (b', k') , k' = i, ∀b' ∈ B';
        Remove from B the blocks in s and those overlapping
        with the blocks in s;
        if q_k is met then
            K^(ℓ) ← K^(ℓ) \ {k'};
        end if
    end if
end for
Phase (K^(c) resource allocation):
repeat
    (b', k') ← arg max_{b∈B, k∈K^(c)} r_{b,k};
    s ← s ∪ (b', k');
    Remove from B the blocks in s and those overlapping
    with the blocks in s;
until B = ∅
```

where $\lceil x \rceil$ denotes the smallest integer bigger or equal to x. 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 M needs to be defined according to the expected throughput per mini-slot as well as its variance. In our numerical results, in Section IV, we have set $M = 10$.

Then, the elements of each category are *re-ordered* in increasing aggregated loss $e_b, b \in \mathcal{B}$. 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 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 $\mathcal{K}^{(\ell)}$

Authorized licensed use limited to: University of Macedonia. Downloaded on July 30,2021 at 11:55:01 UTC from IEEE Xplore. Restrictions apply.

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 (3) and (4) in one go. After each allocation, the allocated blocks are removed from $\mathcal{B}$ and all other categories. This procedure is repeated until the demand of all of the $\mathcal{K}^{(\ell)}$ services are satisfied or no more blocks remain in the categories.

In the last phase of the algorithm, 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 (that have not been allocated to a URLLC service – remember that once a block is allocated it is removed from $\mathcal{B}$). This step is iterated until no more blocks remain available.

The ordering of the utilities has a complexity of $\mathcal{O}\big(\max_{i,k}\{|Cat^i U^k| \log(|Cat^i U^k|)\}\big)$, we conclude that the above is also the overall complexity of the algorithm.

B. NOMA for Downlink Scheduling

In this subsection 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 [12]. In contrast to the scheduling optimization problem as formulated in P0, NOMA allows overlapping amongst the blocks, either full or partial (of some mini-slots). In light of this, P0 is reduced to an analogous linear programming (LP) problem that we refer to as P1, in which the optimization parameter is now a real number $x_{b,k} \in [0,1]$,

$$\text{[P1]} \quad \max_{x_{b,k}\in[0,1]} \sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}^{(c)}} r_{b,k} x_{b,k}, \tag{11}$$

$$\text{s.t.} \quad \sum_{b\in\mathcal{B}} r_{b,k} x_{b,k} \geq q_k, \quad k \in \mathcal{K}^{(\ell)}, \tag{12}$$

$$\sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}} a_{b,i} x_{b,k} \leq \tilde{r}, \quad i \in \mathcal{I}, \tag{13}$$

where $\tilde{r}$ denotes the NOMA (normalized) sum throughput per block ; note that in P0, constraint (4) is upper bounded to unity. This points 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 we simply use $\tilde{r} = 1$.

IV. Numerical Results

In this Section we present results both for the heuristic algorithm in the case of OMA as well as in NOMA.

A. Performance of Heuristic Algorithm

The performance of Algorithm 1 is evaluated for different 5G URLLC configurations and numerologies. To showcase the effectiveness of the proposed heuristic resource allocation algorithm, we compare its performance against the global optimum (achieved through Gurobi optimization solvers) and the LP-LD algorithm discussed in Section II and proposed

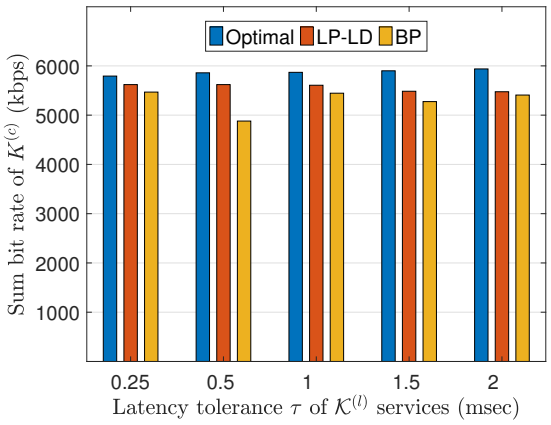

Fig. 2: Sum bit rate of $\mathcal{K}^{(c)}$ services when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 64 kbps. Similar results are produced for demands of 16 and 32 kbps.

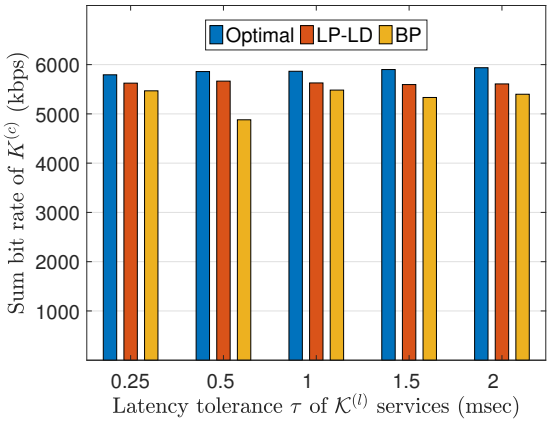

Fig. 3: Sum bit rate of $\mathcal{K}^{(c)}$ services when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 128 kbps.

in [10], using the same simulation setup [1]. This simulation environment was implemented based on the control channel overhead model for supporting the flexible numerology defined in [17] and considers the effect of guard band (i.e., of the cyclic prefix) on the achievable data rate by blocks as modeled in [18].

We measured the bit rates per user in $\mathcal{K}^{(c)}$ for URLLC latency tolerance values $\tau = \{0.25, 0.5, 1, 1.5, 2\}$ msec and bit rate demands $\mathbf{q} = \{16, 32, 64, 128, 256, 512\}$ kbps for five services $k \in \mathcal{K}^{(\ell)}$. The bit rates per user in $k \in \mathcal{K}^{(c)}$ for all the examined algorithms in case of URLLC bit rate demands 16 and 32 kbps are almost same as the ones in case of data demand 64 kbps. Therefore, for the sake of brevity, we omit the presentation of this set of results.

As can be seen through Figs. 2-5, the heuristic algorithm inspired from the reformulation of the scheduling problem as a bin packing optimization, results in a comparable performance

[1]We thank the authors of [10] for kindly sharing their simulation codes in IEEE DataPort.

Authorized licensed use limited to: University of Macedonia. Downloaded on July 30,2021 at 11:55:01 UTC from IEEE Xplore. Restrictions apply.

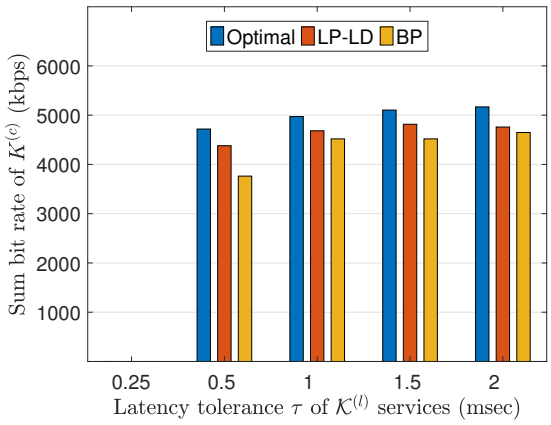

Fig. 4: Sum bit rate of $\mathcal{K}^{(c)}$ services when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 256 kbps.

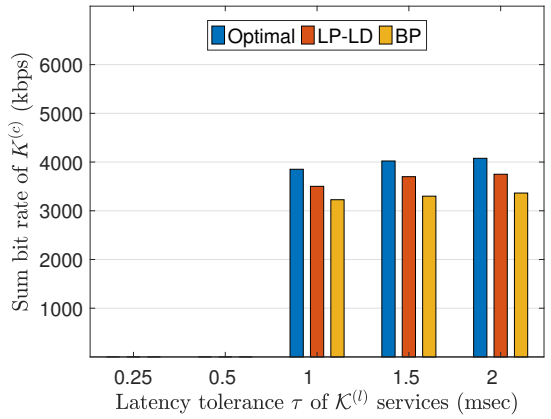

Fig. 5: Sum bit rate of $\mathcal{K}^{(c)}$ services when the bit rate demands of $\mathcal{K}^{(\ell)}$ users are all equal and set to 512 kbps.

to the LP-LD algorithm proposed in [10] and the global optimum (obtained through Gurobi solvers). Note that the proposed algorithm, with no utility computation, no solver used and with near-linear complexity, provides a trade-off between the performance and complexity [2]. 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 (3) and (4) 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.

B. Performance of NOMA

In Figs. 6-9, the sum bit rate for the eMBB services in $\mathcal{K}^{(c)}$ when applying i) NOMA and ii) the optimal OMA scheduling (denoted in Figs. 2-6 by "Optimal" are shown. The

[2]In all executions the processing cost of the LP-LD heuristics was about 20 sec, while that of BP heuristics was about 0.15 sec; a throughout investigation of the processing cost should be considered as a future work.

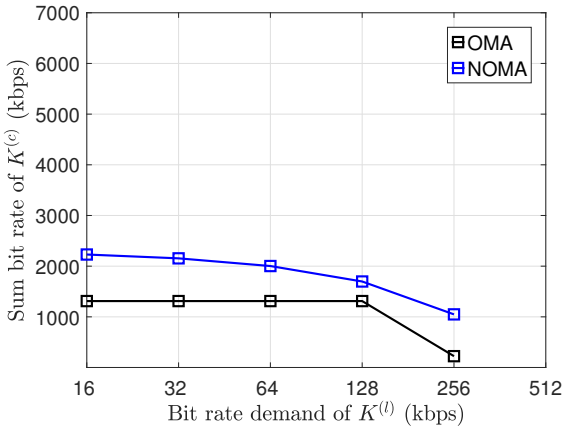

Fig. 6: Sum bit rate for $\mathcal{K}^{(c)}$ services when employing NOMA (blue line) and OMA (black line), for delay tolerance values $\tau = 0.25$. Both schemes result in unfeasible solutions for $q_k = 512$.

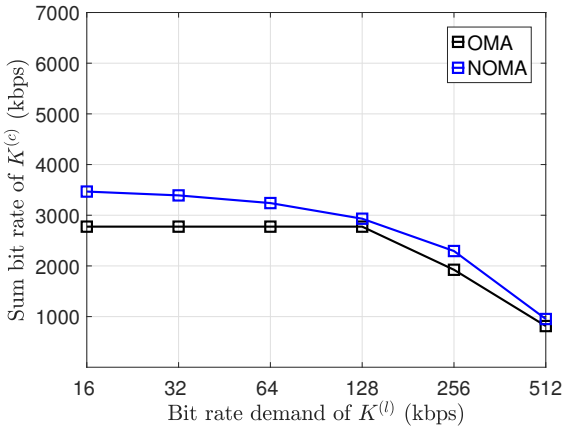

Fig. 7: Sum bit rate for $\mathcal{K}^{(c)}$ services when employing NOMA (blue line) and OMA (black line), for delay tolerance values $\tau = 0.5$.

latency tolerance and bit rate demands considered are $\tau = \{0.25, 0.5, 1, 2\}$ msec and $\mathbf{q} = \{16, 32, 64, 128, 256, 512\}$ kbps. In all cases, as expected, NOMA outperforms the optimal OMA scheduling. The gains are more accentuated at low delay tolerance values, which indicates that NOMA can be beneficial for ultra-low-latency, a scenario of significant practical importance, e.g., in industry 4.0 or vehicle to everything (V2X) communications.

Note that with decreasing the latency tolerance τ_k of services $k \in \mathcal{K}$ a large number of zero throughput mini-slots (void) are generated due to (1). This, decisively reduces the number of available resource blocks, which in turn offers a crucial advantage to the NOMA scheme that allows overlaps.

V. Conclusion

In 5G and beyond, URLLC services will coexist with eMBB services. Layer 2 scheduling in the URLLC and

Authorized licensed use limited to: University of Macedonia. Downloaded on July 30,2021 at 11:55:01 UTC from IEEE Xplore. Restrictions apply.

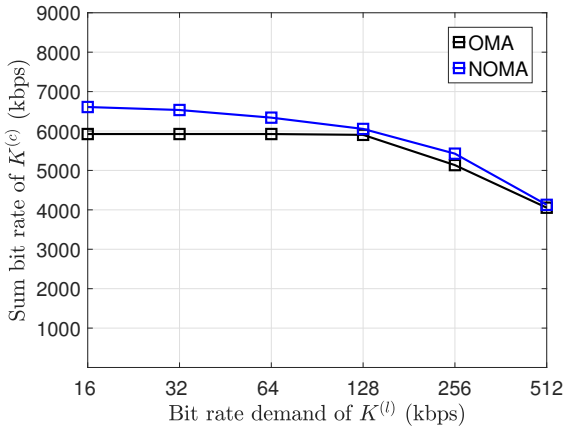

Fig. 8: Sum bit rate for $\mathcal{K}^{(c)}$ services when employing NOMA (blue line) and OMA (black line), for delay tolerance values $\tau = 1$.

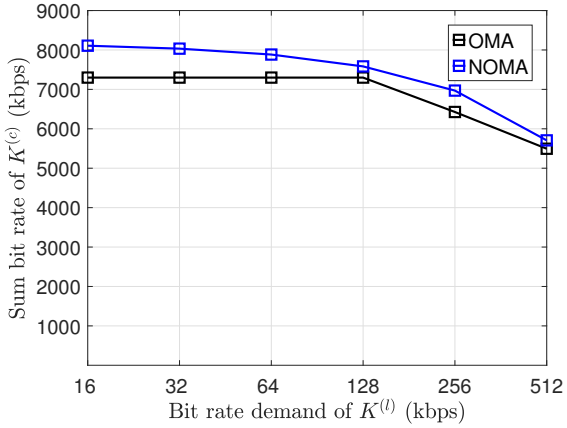

Fig. 9: Sum bit rate for $\mathcal{K}^{(c)}$ services when employing NOMA (blue line) and OMA (black line), for delay tolerance values $\tau = 2$.

eMBB coexistence is a known challenging task. To address this, in this work, we have reformulated the standard eMBB throughput maximization problem as an equivalent conflict minimization, which points to a bin packing setting. Building on this premise, an efficient, near-linear complexity scheduling algorithm was proposed, inspired by a simple greedy heuristic for bin packing problems. In addition to the proposed scheduling using an orthogonal multiple access (OMA) approach, NR also supports non-orthogonal multiple access (NOMA). 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) the proposed near-linear complexity heuristic still provides high resource efficiancy, demonstrating that conflict minimization is indeed key to Layer 2 scheduling, and, ii) there are significant gains in terms of resource utilization when employing NOMA.

REFERENCES

[1] *NR; Physical channels and modulation*, Release 16, Technical Specification (TS) 38.211 V 16.1.0, 3rd Generation Partnership Project (3GPP), 2020.

[2] Y. Sadi, S. Erkucuk, and E. Panayirci, "Flexible physical layer based resource allocation for machine type communications towards 6G," in *Proc. IEEE 2nd 6G Wireless Summit (6G SUMMIT)*, Virtual, Mar. 2020, pp. 1–5.

[3] A. Akhtar and H. Arslan, "Downlink resource allocation and packet scheduling in multi-numerology wireless systems," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshop (WCNCW)*, Barcelona, Spain, Apr. 2018, pp. 362–367.

[4] L. Marijanovic, S. Schwarz, and M. Rupp, "Multi-user resource allocation for low latency communications based on mixed numerology," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Honolulu, Hawaii, USA, Sept. 2019, pp. 1–7.

[5] C. She, C. Yang, and T. Q. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, 2017.

[6] K. Zhang *et al.*, "Dynamic Multiconnectivity Based Joint Scheduling of eMBB and uRLLC in 5G Networks," *IEEE Syst. J.*, early access, Apr. 2020.

[7] 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, 2020.

[8] 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.

[9] 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, 2020.

[10] 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, 2018.

[11] R.-J. Wang, C.-H. Wang, G.-S. Lee, D.-N. Yang, W.-T. Chen, and J.-P. Sheu, "Resource allocation in 5g with noma-based mixed numerology systems," in *IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[12] P. Popovski *et al.*, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[13] 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, 2017.

[14] S. 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, 2016.

[15] A. T. Abusabah and H. Arslan, "NOMA for multinumerology OFDM systems," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–9, 2018.

[16] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, vol 21., 6th ed. Berlin, Germany: Springer-Verlag, 2018.

[17] H. Miao and M. Faerber, "Physical downlink control channel for 5G new radio," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC), Oulu, Finland*, June 2017, pp. 1–5.

[18] 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.

Authorized licensed use limited to: University of Macedonia. Downloaded on July 30,2021 at 11:55:01 UTC from IEEE Xplore. Restrictions apply.