Cost-efficient RAN Slicing for Service Provisioning in 5G/B5G

Cost-efficient RAN Slicing for Service Provisioning in 5G/B5G

Somreeta Pramanik[1], Adlen Ksentini[2], and Carla Fabiana Chiasserini[1]

[1] *Politecnico di Torino, Torino, Italy*
[2] *Communication Systems Department, EURECOM, France*

Abstract

Network slicing represents a substantial technological advance in 5G mobile network, greatly expanding the variety and manifoldness of network services to be supported. Additionally, 3GPP 5G New Radio (NR) has introduced novel features such as mixed numerology and mini-slots, which can be harnessed by network slicing to cater to the diverse requirements of 5G services. While however the co-existence of multiple network slices leads to a challenging resource allocation problem, these new features also severely complicate the management of radio resources. As a further point of attention, the virtualization of radio functions may exact a significant toll from the, already limited, computing resources at the network edge. It follows that a cost-efficient resource allocation across all the slices becomes crucial. In this paper, we address the above-mentioned issues by modeling a cost-efficient radio resource management in 5G NR featuring network slicing, through a Mixed Integer Quadratically constrained Program (MIQCP). We maximize the profit of all slices simultaneously guaranteeing the target data rate and delay specified in the service level agreements (SLAs) fo the different traffic flows. To reduce the complexity of the MIQCP problem, we decompose it into two sub-problems, namely, the scheduling problem of eMBB UEs on a time-slot basis and of uRLLC UEs on a mini-slot basis, while keeping the objective unchanged. To address the scheduling issue of eMBB UEs, we employ a heuristic technique, and, by leveraging the outcome of this heuristic, we derive an optimal solution for the problem of uRLLC UEs. The significance of the proposed approach over a baseline approach is evaluated through extensive numerical simulations in terms of the number of allocated uRLLC RBs per mini-slot. We also assess our approach by measuring the impact of the uRLLC slice changes on the eMBB slice, and vice versa, including delay for uRLLC users and data rates for eMBB users.

1 Introduction

Massive and highly heterogeneous network slicing is a key feature of beyond-5G and 6G networks (B5G/6G), where tenants are not solely focused on vertical industries but are also extending digitalization to the final consumer through new services such as holographic communication, multi-sensory experience, and robotics [1]. In the realm of 6G, networks must effectively handle vast end-to-end slices spanning various technological domains, including radio access network (RAN), edge, cloud, and core, and effectively address the challenges they pose in terms of low-latency communication, high data rate, and increased reliability.

A network slice refers to a virtual network constructed atop a physical network corresponding to a network service, designed to give the slice tenant the perception of operating their dedicated physical network. It provides the flexibility to customize slices, ensuring the fulfillment of various SLAs through the implementation of isolation techniques [2, 3]. Considering the concept of network slicing, 3GPP has classified 5G services into three distinct classes according to their communication service requirement: (i) eMBB, (ii) uRLLC, and (iii) massive Machine Type Communications (mMTC) [4]. In this context, network slicing can help to reduce CAPital EXpenditure (CAPEX)/ OPerating EXpenditure (OPEX) as one physical infrastructure is shared efficiently to fulfill the heterogeneous communication service requirements of emerging network services.

Although network slicing is well-researched, slicing (sharing) the RAN resources is still challenging. Indeed, the new features introduced by 5G NR such as the concept of numerology [5], mini-slot based transmission [6], and punctured scheduling [7] make the management of radio resources more complex. Numerology in 5G NR entails the provision of various frequency domain subcarrier spacings (SCSs) and time domain symbol lengths within the time-frequency orthogonal frequency division multiplexing (OFDM) grid. Numerology flexibility allows for efficient scheduling of eMBB and uRLLC users, selecting SCS and OFDM symbol lengths to meet service requirements.Additionally, the mini-slot approach supports transmission shorter than the regular slot duration. A mini-slot (or the smallest scheduling time unit) occupies 2, 4, or 7 OFDM symbols (regardless of numerology). Finally, the punctured scheduling enables non-orthogonal slicing of radio resources and facilitates the uRLLC traffic to preempt resources that have already been allocated to the eMBB users Taking into account these three techniques, and their potentiality in fulfilling service requirements, makes the RAN slicing a multi-timescale, non-trivial problem.

As an example, using a numerology (μ), the time duration of the physical resource blocks (PRB) is scaled down by a factor 2^{μ} while the frequency is scaled up by 2^{μ}. Thus, using higher numerology and shorter mini-slot duration decreases the RAN latency, but it increases the amount of processing, hence the system energy consumption, since UEs and gNB execute a number of RAN functions 2^{μ} more times per time unit. The trade-off between spectral efficiency and the consumption of data processing resources presents a complex scheduling dilemma. To elaborate, the scarcity of radio spectrum necessitates

efficient spectrum sharing to meet the SLAs of each slice. Simultaneously, the limited computing resources at the network's edge underscore the importance of allocating resources in a computationally aware manner across all slices. Indeed, if a slice's service exhibits elasticity [8], the resource demand of the slice can dynamically change based on the operational computational cost, aiming to maximize the slice's profit. This observation prompts a thorough exploration of the intricate relationship between the cost of computing resources and slice dimensioning.

While the existing state-of-the-art research on 5G NR RAN slicing [9, 10, 11, 12, 13, 14] predominantly concentrates on delivering a satisfactory level of quality of service (QoS) or user's quality of experience (QoE), none of the current studies devises a slicing strategy that is both cost-efficient and considers the real-world interdependence between the cost of computing resources at the network edge and the RAN's capability to support diverse network slices.

To summarize our contributions are as follows.

- We address the challenging problem of cost-efficient resource management in 5G NR featuring network slicing by first formulating it as a *mixed-integer quadratically constrained program* (MIQCP) taking into account (i) different values of numerology, (ii) different mini-slot durations, (iii) different throughput and latency requirements per slice. Our goal is to maximize the expected long-term profit of all slices. Such profit is defined as the difference between the sum of the utility of all eMBB UEs across t time-slots and the normalized cost of computing resource consumption due to the slices supported on the RAN. Importantly, the above problem is NP-hard.

- In light of the problem complexity, we decouple the original problem (P) into two sub-problems, one tackling the resource allocation for eMBB UEs on a time-slot basis (P1), and the other addressing the resource allocation for uRLLC UEs on a mini-slot basis (P2). We then redefine the first sub-problem into a maximization problem for each time slot, and the second sub-problem as a maximization problem for each mini-slot within every time-slot.

- Due to the NP-hardness of P1, we envision a low-complexity heuristic to solve it, thus improving the minimum expected achieved rate (MEAR) among eMBB users (providing the eMBB users with the target). Next, we leverage a M/M/1/k queue to model the delay of the uRLLC users and a utility function for eMBB users to represent the network resources utilization and the target data rate. In so doing, we reformulate P2 taking into account both the decision made by solving P1 and the computing cost associated with the slices. Finally, at every time slot, we solve the new formulation of P2 to maximize the efficiency in resource utilization, while meeting the target eMBB data rate and uRLLC delay.

- We perform a comprehensive experimental analysis for the proposed scheduling approach. We also compare the results in terms of average number

of occupied uRLLC RBs per mini-slot and average delay of uRLLC UEs, against the Static Resource Slicing (SRS) approach [15, 12] where slice requests are processed without considering the CPU cost of the gNB due to slicing. Notice that, to the best of our knowledge, no prior work exists that has developed a cost-efficient/computational-aware RAN slicing strategy for allocating radio resources, allowing for a direct comparison with our proposed CERS approach. More precisely, no prior work has demonstrated the cost-effectiveness in radio resource slicing within the context of 5G NR. We also evaluate the performance of our proposed approach in terms of delay experienced by the uRLLC users and the observed data rates of eMBB users, by measuring the impact that changes occurring in uRLLC slice have on the eMBB slice and vice-versa.

The rest of the paper is organized as follows. Sec. 2 discusses some relevant work while highlighting the novelty of our contribution. Sec. 3 introduces the RAN slicing model and the problem formulation. Sec. 4 describes the proposed solution approach, while Sec. 5 presents pur performance evaluation. Finally, Sec. 6 draws some conclusions and discusses directions for future research.

2 Related Work

Network Slicing has received a great deal of attention owing to its relevance in the support of highly demanding mobile services and applications. In particular, multiplexing between eMBB and uRLLC traffic in a shared RAN has been tackled in [9, 10, 16, 11]. Indeed, given the limited radio resources (e.g., PRBs, transmit power) in a RAN, an efficient resource allocation among eMBB and uRLLC slices is crucial to satisfy the QoS requirements of the users. To facilitate the support of the slices, 5G NR standardized the techniques of numerology [17], mini-slot based transmission [6], and punctured scheduling [7] to be used for service multiplexing in a RAN. Taking into account these three techniques, the RAN slicing has become a multi-timescale problem.

The existing body of work can be categorized into two main lines of research. The former pertains to the orthogonal slicing approach, where the wireless service provider reserves a portion of bandwidth for the eMBB users, and another portion of bandwidth for the uRLLC users. In this approach, which is considered for instance in [18, 19, 20, 21, 22], service isolation among network slices is provided. However, the allocated resources to uRLLC slice may be underutilized due to the uRLLC traffic dynamics. Conversely, the latter line of research uses non-orthogonal slicing with punctured scheduling. This approach, which is used in [9, 10, 11, 12, 13, 14, 23], can provide an efficient use of radio resources for uRLLC users. However, punctured scheduling may degrade the performance of eMBB slice due to the potential reduction of the eMBB users' data rate.

More in details, an example of the first approach can be found in [24] where we designed a cost-efficient slicing strategy, named CES, that minimizes the computing cost due to slicing, while guaranteeing the target data rate for eMBB users and delay of uRLLC users specified in the SLA. Looking at the second

approach, instead, Bairagi et al. [9] considered the network slicing problem in a downlink orthogonal frequency division multiple access (OFDMA) system by maximizing the spectral efficiency, while guaranteeing the required data rate for the eMBB users and latency for uRLLC users, based upon puncturing technique. Anand et al. [10] considered a joint eMBB/uRLLC scheduling problem for various eMBB rate loss models while the uRLLC traffic is dynamically multiplexed with the eMBB traffic through punctured scheduling. Alsenwi et al. [11] proposed a risk-sensitive punctured scheduling approach, where the radio resources used by the eMBB users can be reallocated to the uRLLC users. Also, [12] proposed Mixed numerology Mini-slot based Resource Allocation [MiMRA] that guarantees that the loss in eMBB data rate due to the co-existing uRLLC traffic is minimal. The work in [13], instead, aims to maximize the minimum expected achieved rate of eMBB users (MEAR), and fairness among them, by employing a one-to-one matching game to compute appropriate eMBB and uRLLC pairs for uRLLC resource allocation. Finally, [25] studied the resource slicing problem and formulated it as an optimization problem that aims at maximizing the eMBB data rate subject to a uRLLC reliability constraint, while accounting for the variance of the eMBB data rate to reduce the impact of immediately scheduled uRLLC traffic on the eMBB reliability.

Novelty. Compared to the works presented above, in this paper we apply a non-orthogonal slicing approach with punctured scheduling that accounts for both the transmission priority of the uRLLC traffic and its dynamics, and, even more importantly, the computational cost of such non-orthogonal slicing. Specifically, we study the radio resource slicing problem for serving eMBB and uRLLC users in a downlink OFDMA-based RAN by leveraging numerology and punctured scheduling through mini-slot based transmission to serve uRLLC users. It is worth noting that, although some of the recent related work, such as [12, 9, 13, 26, 27], address the technical challenges in the eMBB and uRLLC co-existence problem, no existing work considers both the co-existence problem and cost-efficient slicing strategies. Instead, by a tractable methodology, we are able to address, and effectively reduce, the computing cost due to slicing with respect to traditional approaches while guaranteeing the target data rate of eMBB users and delay of uRLLC users specified in the SLA.

Table 1: Parameters of different 5G numerology settings [26]

Numerology	0	1	2	3
Subcarrier spacing (SCS)	15 kHz	30 kHz	60 kHz	120 kHz
PRB bandwidth	180 kHz	360 kHz	720 kHz	1.44 MHz
Time slot duration	1 ms	0.5 ms	0.25 ms	0.125 ms

Table 2: 5G numerology and the considered uRLLC transmission duration

μ	uRLLC transmission duration	Blocklength per PRB
0	2	24
1	4	48
2	8	96

Table 3: Summary of notations

Symbol	Meaning
$\mathcal{S}$	Set of slices
$\mathcal{E}$	Set of eMBB users
$\mathcal{U}$	Set of uRLLC users
$\mathcal{F}$	Set of RBs of uniform bandwidth B
$\mathbf{U}$	Set of numerologies
B_μ	Bandwidth of an RB in numerology μ
τ_μ	Duration of a time-slot in numerology μ
ω	Duration of a mini-slot
$\mathcal{M}$	No. of mini-slots in a time-slot
$\mathcal{T}$	Total number of time-slots
$R_u^{m,t}$	Achieved data rate of an uRLLC user at mini-slot m of time-slot t
r_e^t	PRB rate for eMBB user e
R_e^t	Achieved data rate of an eMBB user at time-slot t
R_{min}	Minimum expected achieved rate (MEAR) among all eMBB users
U	Utility function for an eMBB user
ϕ_s	CPU cost function for deploying slice s
λ	arrival rate of uRLLC traffic in a mini-slot m of time-slot t
D_{max}	Maximum tolerable uRLLC delay
x_{th}	Target data rate of an eMBB user
α	Resource allocation vector for an eMBB user
ζ	Resource allocation vector for punctured eMBB and uRLLC pairs
C	Constant number of RBs
γ_e^t	SNR of eMBB user e in time-slot t
$\gamma_u^{m,t}$	SNR for uRLLC user u from gNB at mini-slot m of time-slot t

3 System Model and Problem Formulation

For simplicity, we start by considering a scenario with one gNB serving two user groups: $\mathcal{E}$, which requires eMBB service, and $\mathcal{U}$, which demands uRLLC service. In our simplified notation, we have a set of slices $\mathcal{S}$, consisting of a single eMBB slice and a single uRLLC slice, although the extension to multiple eMBB and uRLLC slices is straightforward. Radio resources in the frequency domain are divided into RBs $j \in \mathcal{F} = \{1, 2, 3, ..F\}$, each with a bandwidth B determined by

the numerology (μ) chosen (as shown in Table 1). The time domain is divided into time slots $\mathcal{T} = \{1, \dots t\}$, each with a duration τ depending on μ. These time slots are further subdivided into mini-slots $\mathcal{M} = \{1, \dots m\}$, with each mini-slot duration ω calculated based on the number of OFDM symbols. The arrival of uRLLC traffic at the gNB follows a Poisson distribution and occurs during any mini-slot m of a given time slot t. Each uRLLC UE $u \in \mathcal{U}$ requests a payload of size $L_u^{m,t}$ (varying from 32 to 200 bytes). gNB allots the RBs to the eMBB UEs at the commencement of any time slot $t \in \mathcal{T}$.

The achievable data rate of an uRLLC user among overlapped RBs when multiple RBs are allocated at a mini-slot m of time-slot t is given as:

$$R_u^{m,t} = \sum_{j \in \mathcal{F}} \sum_{e \in \mathcal{E}} \zeta_{e,u,j}^{m,t} \cdot r_{u,j}^{m,t} \,. \tag{1}$$

where $\zeta_{e,u,j}^{m,t}$=1 indicates that $j \in \mathcal{F}$ RB of eMBB UE $e \in \mathcal{E}$ pairs with an URLLC user $u \in \mathcal{U}$ using puncturing at a mini-slot $m \in \mathcal{M}$ of time-slot $t \in \mathcal{T}$, and $\zeta_{e,u,j}^{m,t}$=0 otherwise. $r_{u,j}^{m,t}$ is the achievable rate of an RB j of an uRLLC user u. The data rate falls in the finite block length channel coding regime due to short-sized packet transmission of uRLLC and is approximated as, [28]

$$r_{u,j}^{m,t} = B_\mu \log_2(1 + \gamma_{u,j}^{m,t}) - \sqrt{\frac{C_{u,j}^{m,t}}{l_{u,j}^{m,t}}} Q^{-1}(\epsilon) \log_2 e) \tag{2}$$

where $l_{u,t}^{m,t}$ represents the length of the codeword block in symbols and can be obtained according to Table 2 based on the selected μ for the uRLLC slice. $\gamma_{u,j}^{m,t}$ is the signal-to-noise ratio (SNR) of UE u, $C_{u,j}^{m,t}$ is the channel dispersion, representing the stochastic variability of the channel compared to a deterministic channel with the same capacity, given by $C_{u,j}^{m,t} = 1 - \frac{1}{(1+\gamma_{u,j}^{m,t})^2}$, $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function, ϵ is the transmission error probability.

For conventional services, such as eMBB with large transmitted packet size, the achievable data rate of an eMBB user e for a given RB at time slot t can be directly estimated according to Shannon's capacity as,

$$r_{e,j}^t = B_\mu \log_2(1 + \gamma_{e,j}^t) \tag{3}$$

where $\gamma_{e,j}^t = \frac{P_e \cdot |h_{e,j}^t|^2}{N_e}$ represents the SNR. P_e, h_e, and N_e indicate the transmission power, channel gain, and channel noise, respectively, for user $e \in \mathcal{E}$. The achievable rate of the eMBB UE, $e \in \mathcal{E}$, in Transmission Time Interval (TTI) t is given by:

$$R_e^t = \left\{ \sum_{j=1}^{F} \alpha_{e,j}^t - \sum_{j=1}^{F} \sum_{u \in \mathcal{U}} \sum_{m \in \mathcal{M}} \zeta_{e,u,j}^{m,t} \right\} \cdot r_{e,j}^t \tag{4}$$

where binary variable $\alpha_{e,j}^t = 1$ indicates that the j-th RB is allocated to UE e at TTI t, and $\alpha_{e,j}^t = 0$ otherwise,

The average achievable data rate for the eMBB user $e \in \mathcal{E}$ is then given by,

$$\overline{R}_e = \frac{1}{|T|} \sum_{t=1}^{|T|} R_e^t \,. \tag{5}$$

Crucially, the eMBB data rate loss is linked to the overlapping technique (puncturing) of uRLLC. Thus, eMBB users that lose their resources by sharing their allocated resources with uRLLC users should be guaranteed a more significant proportion of resources in the long run. We therefore consider as primary performance metric for eMBB users the Minimum Expected Achieved Rate (MEAR) [13, 9], i.e.,

$$R_{\min} = \min_{e \in \mathcal{E}} (\overline{R}_e) \,. \tag{6}$$

Next, we introduce the SLA model, which includes both data rate and packet latency as performance metrics. While the former can be derived by aggregating the amount of data that is successfully transmitted over time, a queuing model of UEs' packets is needed to derive the latter. To this end, we assume that each uRLLC slice has its down-link queue at the gNB, and all packets belonging to a slice share the same queue. We then model the uRLLC slice queue at the gNB as an M/M/1/K queue with service rate σ and traffic arrival rate λ [22]. As σ depends upon the scheduling process at the MAC layer, while λ corresponds to the traffic rate of the users running on top of the slice, we write:

$$\sigma_{u,m,t} = \frac{\sum_{j \in \mathcal{F}} \zeta_{e,u,j}^{m,t} \cdot R_u^{m,t}}{L} \tag{7}$$

$$\lambda = \frac{|\mathcal{U}| \cdot d_{u,m,t}}{L} \tag{8}$$

where L is the packet size of the uRLLC application, $|\mathcal{U}|$ is the number of UEs belonging to the uRLLC slice, $d_{u,m,t}$ is the traffic arrival rate of uRLLC service per user in each mini-slot m of time-slot t. The average number of customers in an M/M/1/K system is:

$$q_{u,m,t} = \frac{1 - \rho_{u,m,t}}{1 - \rho_{u,m,t}^{K+1}} \sum_{k=0}^{K} k \rho_{u,m,t}^k \tag{9}$$

where $\rho_{u,m,t} = \frac{\lambda}{\sigma_{u,m,t}}$. The average number of customers waiting in the queue is:

$$L_{u,m,t} = q_{u,m,t} - (1 - p_0) \tag{10}$$

Little's law can then be applied to estimate the latency experienced by uRLLC packets in the corresponding queue:

$$\delta_{u,m,t} = \frac{L_{u,m,t}}{\overline{\lambda}} \tag{11}$$

where, $\overline{\lambda} = \sum_{n=0}^{K-1} \lambda * p_n$, p_n is the probability of n customers in the system. At mini-slot m of time-slot t, the delay of a packet arriving at the u-th UE is given by the sum of transmission delay and queuing delay,

$$D_{u,m,t} = W_{u,m,t} + \delta_{u,m,t} \tag{12}$$

where the transmission delay, $W_{u,m,t}$, is the queue service time, which depends upon the data-rate used to transmit towards the UE (see (2)).

3.1 The Cost-Effective RAN Slicing (CERS) Strategy

Our objective is to derive an optimal RAN slicing control strategy in 5G NR that maximizes the long-term profit of all slices. This profit is defined as the difference between the utility of eMBB UEs across t time-slots and the normalized cost attributed to the computational resource consumption arising from the supported slices on the RAN. The utility of eMBB users is given by:

$$U = \begin{cases} 1 - \text{erf}(x^{th} - x^o) & \text{if } x^o \geq x^{th} \\ \text{erf}(x^{th} - x^o) & \text{otherwise} \end{cases} \tag{13}$$

where x^{th} is the target per-UE data rate for eMBB traffic and x^o is the observed minimum expected achieved data rate (MEAR) over all eMBB users (i.e., the observed value of $R_{\min}$). To meet SLAs, in this case, the observed data rate, it is crucial to allocate radio resources so that the observed data rate consistently meets or stays below target values (thresholds). Moreover, it is crucial to maintain the observed data rate as close as possible to the respective target, avoiding overshooting it for optimal utilization of network resources. Therefore, our selection of the utility function takes into account these essential properties.

The computing resource consumption for deploying slice $s \in \mathcal{S}$, denoted with ϕ_s, is instead based on our experimental findings [29, 30] and is given by:

$$\phi_s = 3.9 \cdot n_s + 0.44 \cdot a_s + 30 \quad \forall s \in \mathcal{S} \tag{14}$$

where n_s is the number of users served by slice s and a_s is the number of RBs allocated to the slice.

By taking $\alpha_{e,j}^t$ and $\zeta_{e,u,j}^{m,t}$, indicating the RBs allocation for the eMBB and uRLLC slices (resp.), as decision variables, the CERS problem formulation can

then be written as:

$$\mathbf{P_0}: \max_{\{\alpha\},\{\zeta\}} U(\{\alpha\},\{\zeta\}) - \mathbb{E}_{t\in\mathcal{T}}\left[\sum_{s\in\mathcal{S}} \phi_s^t(\{\alpha\},\{\zeta\})\right] \tag{15}$$

$$\text{s.t. } D_{u,m,t} \leq D_{max}, \quad \forall u \in \mathcal{U}, m \in \mathcal{M}, t \in \mathcal{T} \tag{15a}$$

$$\sum_{e\in\mathcal{E}} \alpha_{e,j}^t \leq 1, \quad \forall j \in \mathcal{F}, \forall t \in \mathcal{T} \tag{15b}$$

$$\sum_{e\in\mathcal{E}} \sum_{u\in\mathcal{U}} \zeta_{e,u,j}^{m,t} \leq 1, \quad \forall j \in \mathcal{F}, m \in \mathcal{M}, t \in \mathcal{T} \tag{15c}$$

$$\sum_{j\in\mathcal{F}} \sum_{e\in\mathcal{E}} \alpha_{e,j}^t \leq |\mathcal{F}|, \forall t \in \mathcal{T} \tag{15d}$$

$$\sum_{j\in\mathcal{F}} \sum_{e\in\mathcal{E}} \sum_{u\in\mathcal{U}} \zeta_{e,u,j}^{m,t} \leq |\mathcal{F}|, \forall m \in \mathcal{M}, t \in \mathcal{T}, \tag{15e}$$

$$\sum_{j\in\mathcal{F}} \zeta_{e,u,j}^{m,t} \geq 1, \forall e \in \mathcal{E}, u \in \mathcal{U}, m \in \mathcal{M}, t \in \mathcal{T} \tag{15f}$$

$$\alpha_{e,j}^t, \zeta_{e,u,j}^{m,t} \in \{0,1\} \quad , \forall u \in \mathcal{U}, e \in \mathcal{E}, j \in \mathcal{F}\, m \in \mathcal{M}, t \in \mathcal{T} \tag{15g}$$

where, for clarity, in the objective function we highlighted the dependency of the utility and of the computational cost of a slice on the number of radio resources ($\{\alpha\}$ and $\{\zeta\}$) allocated to eMBB and uRLLC users (resp.).

The uRLLC latency constraint is established in (15a), which guarantees that the uRLLC users' packet delay will not exceed the target value D_{max}. Constraint (15b) states that every RB can be allocated to at most one eMBB user, while (15c) ensures that every RB is used by at most one uRLLC user. The total number of resources allocated to all eMBB users in the system is constrained by (15d). Additionally, (15e) places a limit on the maximum number of RBs that can be allocated to arriving uRLLc users within a mini-slot. Constraint (15f) guarantees the allocation of at least one RB to a uRLLC user. Constraint (15g) specifies that each vector element of α, ζ is binary.

The problem formulation, along with the constraints, results in a mixed-integer quadratically constrained problem (MIQCP), which is NP-hard. It is thus essential to simplify the problem to reduce its computational complexity and make it solvable in a reasonable time in practical system scenarios.

3.2 Proof

The problem (P_0) can be proved to be NP-hard by using a reduction from the knapsack problem (P_K), a combinatorial optimization problem, known to be NP-hard.

Definition (P_0): Cost-efficient radio resource slicing in 5GNR involves allocating the finite radio spectrum into multiple slices in a cost-efficient manner to meet diverse and conflicting service requirements within the constraints of limited resources and a target delay.

Known NP-Hard problem (P_K): In the classic knapsack problem, the objective is to select items, each with a given weight and value, to maximize the total value without surpassing a weight limit.
Reduction Mechanism: To effectuate this reduction, we conceptualize the knapsack's capacity constraints as analogous to the bandwidth (Total number of available RBs) and timing constraints (a target delay) in P_0. The items in P_K, with their respective weights and values, are paralleled to the number of RBs allocated per slice in P_0, where the objective morphs into maximizing the profit of slices within the predefined constraints of bandwidth and timing.
Contradiction Argument: If P_0 is solvable in polynomial time (i.e., not NP-hard), then so would be P_K, contradicting P_K's NP-hardness.
Conclusion: The polynomial-time reducibility of P_K to P_0 implies P_0 is also NP-hard, as solving it efficiently would inadvertently solve the knapsack problem, an NP-hard problem, efficiently.

4 Optimization Method

In light of the complexity of the optimization problem $\mathbf{P_0}$, we envision a lower-complexity solution strategy by leveraging the concept of divide-and-conquer [31]. We thus divide $\mathbf{P_0}$ into two sub-optimization problems and solve the new problems as set forth below:

- Subproblem 1 ($\mathbf{P_1}$) – Resource allocation for eMBB UEs on a time-slot basis
- Subproblem 2 ($\mathbf{P_2}$) – Resource allocation for uRLLC UEs on a mini-slot basis.

Subproblem 1. Given the short duration of a time slot, it is fair to assumed that eMBB UEs have a high demand for data over the whole considered slot. Consequently, at the beginning of every time slot, $t \in \mathcal{T}$, the eMBB users are allocated with RBs, and the allocated resources remain unchanged throughout the time slot. Then, by setting, in this first stage, all $\zeta_{e,u,j}^{m,t}$'s equal to zero, we formulate the first sub-problem as:

$$\mathbf{P_1}: \max_{\{\alpha\}} U(\{\alpha\}) - \mathbb{E}_{t\in\mathcal{T}}\left[\sum_{e\in\mathcal{E}} \phi_e^t(\{\alpha\})\right] \tag{16}$$

$$\text{s.t.} \quad \sum_{e\in\mathcal{E}} \alpha_{e,j}^t \leq 1, \quad \forall j \in \mathcal{F}, \forall t \in \mathcal{T} \tag{16a}$$

$$\sum_{j\in\mathcal{F}}\sum_{e\in\mathcal{E}} \alpha_{e,j}^t \leq |\mathcal{F}|, \quad \forall t \in \mathcal{T} \tag{16b}$$

$$\alpha_{e,j}^t \in \{0,1\}, \quad \forall e \in \mathcal{E}, j \in \mathcal{F}\, t \in \mathcal{T}. \tag{16c}$$

Subproblem 2. When uRLLC traffic requests arrive during any mini-slot m of time slot t, the scheduler aims to fulfill these requests in the subsequent

mini-slot $(m+1)$. The task involves evaluating suitable eMBB users to pair with the set of arrived uRLLC users while maintaining fairness among eMBB users. We then set in $\mathbf{P_0}$ all $\alpha^t_{e,j}$'s to the values obtained by solving $\mathbf{P_1}$, and we formulate the second sub-problem as follows:

$$\mathbf{P_2}: \max_{\{\zeta\}} U(\{\zeta\}) - \mathbb{E}_{t\in\mathcal{T}}\left[\sum_{u\in\mathcal{U}} \phi^t_u(\{\zeta\})\right] \tag{17}$$

$$\text{s.t. } D_{u,m,t} \leq D_{max}, \quad \forall u \in \mathcal{U}, m \in \mathcal{M}, t \in \mathcal{T} \tag{17a}$$

$$\sum_{e\in\mathcal{E}}\sum_{u\in\mathcal{U}} \zeta^{m,t}_{e,u,j} \leq 1, \quad \forall j \in \mathcal{F}, m \in \mathcal{M}, t \in \mathcal{T} \tag{17b}$$

$$\sum_{j\in\mathcal{F}}\sum_{e\in\mathcal{E}}\sum_{u\in\mathcal{U}} \zeta^{m,t}_{e,u,j} \leq |\mathcal{F}|, \quad \forall m \in \mathcal{M}, t \in \mathcal{T}. \tag{17c}$$

To further clarify the above solution approach, let us refer to the following simple example. Consider that, at the beginning of time slot $t-1$, there are 3 eMBB UEs and each is assigned 4 RBs. Within $t-1$, a service request for uRLLC UEs arrives and the necessary RBs are allocated as overlapped uRLLC traffic in the mini-slots. For instance, during this time, 4, 7, and 2 RBs of eMBB UEs 1, 2, and 3 are allocated to uRLLC UEs, respectively. Therefore, the data rate of eMBB UEs 1, 2, and 3 drops by 4 RBs·1 mini-slot, 7 RBs·1 mini-slot, and 2 RBs·1 mini-slot, respectively. At the start of the next time slot, t, the gNB acknowledges the resource scheduling of uRLLC UEs in time slot $t-1$ to compensate eMBB UE 1, 2, and 3 for their reduced data rate. In particular, the gNB will allocate more RBs to such eMBB users in a fair manner, that is, with, e.g., UE 2 receiving a higher number of additional allocated RBs than 3.

4.1 Low-Complexity Heuristic for Sub-Problem $\mathbf{P_1}$

To ensure a fair share of resources among the eMBB users, resource allocation at a given time slot t has to account for the data rate such users experienced in the previous time slot $(t-1)$. As $\mathbf{P_1}$ (16) is still an NP-hard problem, a low-complexity resource allocation algorithm has to be used. To this end, we draw on the solution proposed in [9, 13] and enhance it to adapt it to our specific problem. The algorithm we apply consists of the following steps:

1. Initialization: A fixed number of RBs, N, are initially allocated to every eMBB user $e \in \mathcal{E}$, so that the target eMBB data rate is fulfilled.

2. At the beginning of slot $t \in T$, evaluate previously achieved data rates of all eMBB users. That is, get $R^{t-1}_e, \forall e \in \mathcal{E}$ from eMBB-uRLLC pairing and uRLLC resource allocation by solving $\mathbf{P_2}$.

3. For each RB $j \in \mathcal{J}$, with $|\mathcal{J}| = |\mathcal{F}| - N \cdot |\mathcal{E}|$, compute the rationality factor for every eMBB user e, defined as

$$H(e) = \frac{R^t_e + R^{t-1}_e}{\overline{R}^{t-1}}. \tag{18}$$

4. Assign RB $j \in \mathcal{J}$ to the user with the least value of $E(e)$.

5. Repeat step 3 to 4 for all the available RBs in $\mathcal{J}$.

Algorithm 1 Heuristic Algorithm for Solving $\mathbf{P_1}$

Input: R_e^{t-1} for each e in $\mathcal{E}$.
Output: $E(idx)$-updated resource allocation for each e in $\mathcal{E}$.

Initialize:
  $i = 1$, $E(e) \leftarrow N, \forall e \in \mathcal{E}$
while $i \leq |\mathcal{T}|$ **do**
  Get $R_e^{t-1}, \forall e \in \mathcal{E}$ from eMBB-uRLLC pairing and uRLLC resource allocation (Solve $\mathbf{P_2}$).
  $remRB \leftarrow |\mathcal{F}| - N \cdot |\mathcal{E}|$
  for $j = 1 : remRB$ **do**
    for $k = 1 : |\mathcal{E}|$ **do**
      $H(k) \leftarrow \frac{nRB \cdot r_k^t + R_e^{t-1}}{\overline{R}^{t-1}}$
    end for
    $idx \leftarrow \{e : e = \text{argmin}_{\mathcal{E}} H(e)\}$
    $E(idx) \leftarrow E(idx) + 1$
  end for
  $i \leftarrow i + 1$
end while

To summarize, at t=1, the algorithm allocates resources equally (i.e., N RBs to each eMBB user). Then, it allocates resources to eMBB UEs in the rest of the time slots depending on the previous time slot. More specifically, it considers the rationality $H(e)$, which is the fraction of the sum of achieved data rate of a given eMBB user involving the current time-slot t ($R_e^t = N \cdot r_e^t$) and the previous time slot ($t-1$) (R_e^{t-1}) relative to the average achieved data rate across all eMBB users ($\overline{R}^{t-1}$). A low achieved eMBB data rate in the previous time slot results in a lower rationality for a particular eMBB user. Thus, the eMBB user with the least achieved data rate due to uRLLC puncturing of eMBB RBs or weak channel conditions in time slot $t-1$ has higher priority to be allocated the RB. In this way, the algorithm can accommodate the MEAR of eMBB UEs in the long run adequately and in a fair manner.

Complexity analysis. For each time slot, let N be the number of RBs assigned initially to all eMBB UE where $N \leq |\mathcal{F}|$. The remaining number of RBs to be allocated to the most suffering eMBB users is $(|\mathcal{F}| - N)$. The complexity required for each RB allocation is $O(|\mathcal{E}|)$. The eMBB resource allocation in each time slot takes $((|\mathcal{F}| - N)|\mathcal{E}|)$. It follows that the overall complexity is $O(|\mathcal{T}|\mathcal{F}||\mathcal{E}|)$.

4.2 Solving Sub-Problem $\mathbf{P_2}$

We reformulate the second sub-problem (17) to take into account the CPU cost associated with both the eMBB and the uRLLC slice. Thus, we write $\mathbf{P_2}$ as:

$$\max_{\{\zeta\}} \sum_{e \in \mathcal{E}} U(\{\zeta\}) - \mathbb{E}_{t \in \mathcal{T}} \left[\sum_{s \in \mathcal{S}} \phi_s^t(\{\zeta\}) \right]. \tag{19}$$

In contrast to the definition of U in Eq. (13), x^o here is the average achievable data rate of an eMBB user e in time-slot t and is given by Eq. (4).

Complexity analysis. The problem formulation (19), along with the constraints (17a–17c), results in an MIQCP problem, which can be solved using Gurobi [32]. To solve the model, the non-linear functions (objective function and quadratic constraints) are approximated as piece-wise linear functions. Then, a feasible solution is found, either by a MIP heuristic or by branching. When the gap between the best feasible solution and the best bound is smaller than the default MIPGap parameter (set to 10^{-4}), it is considered that the optimal solution has been attained.

5 Numerical Analysis

In this section, we first describe the scenario we use for our performance evaluation. Then we show the performance of our proposed approach, CERS, through an extensive experimental analysis, and compare it against the Static Resource Slicing (SRS) approach [15, 12] where slice requests are processed without considering the CPU cost of the gNB due to slicing. As mentioned, SRS has been selected as benchmark, since, to the best of our knowledge, no existing scheme for radio resource allocation accounts for cost-efficient/computational-aware RAN slicing.

Table 4: Simulation parameters

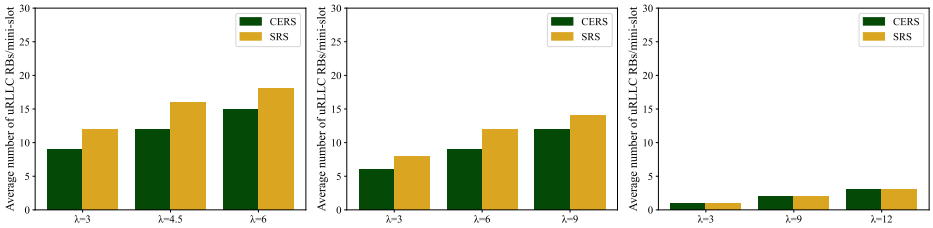
Parameter	Value
Total channel bandwidth	20 MHz
Carrier frequency	2.62 GHz
Maximum BS transmission power	24 dBm
BS coverage radius	500 m
Noise spectral density	-114 dBm
Channel Model	FSPL
Numerology(μ)	{0, 1, 2}
Mini-slot duration	0.25 ms
Target uRLLC delay	1 ms
uRLLC packet size (bytes)	32
Number of UEs	6 (eMBB (3), uRLLC (3))

5.1 Reference Scenario

In our study, we consider a shared 5G NR infrastructure with coexisting uRLLC and eMBB users. We consider one gNB operating in the Frequency Range (FR)-1, with a maximum transmission power of 24 dBm and covering a radius of 500 m. The transmission occurs at the 2.5 GHz frequency band with a total channel bandwidth of 20 MHz. The arrival of uRLLC traffic at mini-slot m of time-slot t follows a Poisson distribution with mean λ, and the uRLLC packet size is set to 32 bytes. We adopt a full buffer model for eMBB buffers at the base station, assuming a continuous data flow. The gNB utilizes numerology $\mu = 0, 1, 2$ to transmit eMBB and uRLLC traffic over all of the available RBs in each numerology. The corresponding time slots for each numerology, $t_{\mu=0} =$ 1 ms, $t_{\mu=1} = 0.5$ ms, and $t_{\mu=1} = 0.25$, are sub-divided into a number of M_0, M_1, and M_2 mini-slots, respectively. The mini-slot duration (ω) is 250 μs, which is sufficient to meet the latency requirement for uRLLC traffic, and it is kept the same for all considered numerologies. Additionally, the simulation incorporates a maximum tolerable delay of 1 ms, with the consideration that eMBB traffic is not as time-sensitive as uRLLC traffic.

5.2 CERS Performance Evaluation

We showcase the effectiveness of our proposed slice-dimensioning method, CERS, taking into account the performance requirements of the eMBB and uRLLC slices. We configure the system parameters as outlined in Table 4, and we compare the slice profit of CERS to static resource slicing (SRS), as shown in Fig. 1, where minimizing the number of RBs assigned to a slice leads to higher slice profit. In SRS, slice requests are processed without considering the CPU cost of the gNB due to slicing. The objective of the SRS scheduler (similar to the Sum-Rate [15] scheduler, MiMRA [12]) is to maximize the average sum rate of eMBB users using the puncturing strategy. In our analysis, we consider two distinct slices namely, eMBB and uRLLC.



(a) Numerology μ=0 (b) Numerology μ=1 (c) Numerology μ=2

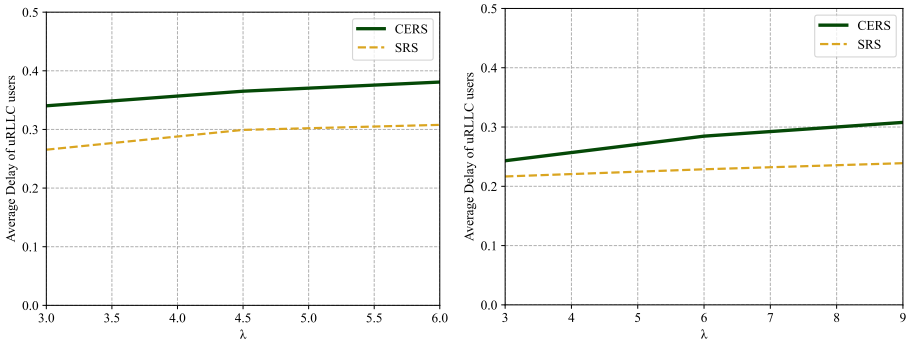
Figure 1: Comparison of the number of RBs allocated to the uRLLC slice at each mini-slot of a time-slot, under CERS and SRS for different uRLLC traffic demands (λ). The traffic demand of each eMBB UE is set to 4 Mbps and the numerology (μ) considered is 0, 1, 2 in (a) (b), and (c), respectively.

Comparison of slice profit. Fig. 1 represents the number of RBs allocated to the uRLLC slice every mini-slot, under our proposed scheme (CERS) and under the considered benchmark (SRS). The target data rate of every eMBB user is set to 4 Mbps and the traffic demand (λ) of every uRLLC user is varied. The results depict that the number of RBs allocated to uRLLC users in every mini-slot is always lower under CERS compared to SRS for every uRLLC traffic demand in Numerology 0 (1a) and 1 (1b), while it is the same for Numerology 2 (1c). CERS indeed maximizes the slice profit by allocating a lower number of RBs than SRS while satisfying the SLAs: the higher the number of RBs allocated to a slice, the higher the CPU cost/utilization of the RAN due to slicing of the radio resources, and the lower the slice profit.

To further illustrate the comparison based on the numerology schemes, it is worth mentioning that in higher numerology schemes (e.g., μ=1 and μ=2) the number of allocated uRLLC RBs is noticeably less compared to lower numerology scheme (μ=0). This reduction is due to the higher PRB rate, scaled up by a factor of 2^{μ}, in the higher numerology schemes. For instance, when the traffic demand λ is set to 3, the allocated RBs in μ=1 and μ=2 are significantly fewer than those in μ=0. Building on our earlier discussion regarding our proposed cost-efficient scheme (CERS), it becomes evident that the impact on CPU cost/consumption increases with the rising number of required RBs. In the case of Numerology 2 (1c), where the required RBs are fewer, CERS experiences a reduced impact on CPU cost, ultimately resulting in the number of allocated RBs being equivalent to that of SRS. To showcase/demonstrate the effectiveness of our proposed approach, specifically in terms of the number of allocated RBs, in Fig. 1b and Fig. 1c, we deliberately select higher values of traffic demand (λ). We remark that we evaluated the performance of CERS only in terms of the number of allocated radio resources, since, as it can be noted in our earlier work [29, 30], the dominant impact on the CPU consumption is represented by the number of connected UEs, rather than by the number of allocated RBs. In addition, we would like to highlight that further considerations about the CPU consumption can be made starting from the results in Fig. 1a and Fig. 1b, which show how CERS allocates a lower number of RBs, with respect to SRS. The smaller the number of radio resources allocated, the lower the CPU utilization of the virtual gNB according to Eq. 3 in [30].

eMBB and uRLLC Slice Performance. The performance of slices can be effectively evaluated by gauging the influence that alterations, such as shifts in traffic demand, in one slice exert on another. In our assessment of the proposed approach CERS, we focused on measuring performance in terms of the delay encountered by uRLLC users and the observed data rates of eMBB users. This evaluation was conducted by varying the uRLLC traffic demand and adjusting the target data rate of eMBB users.

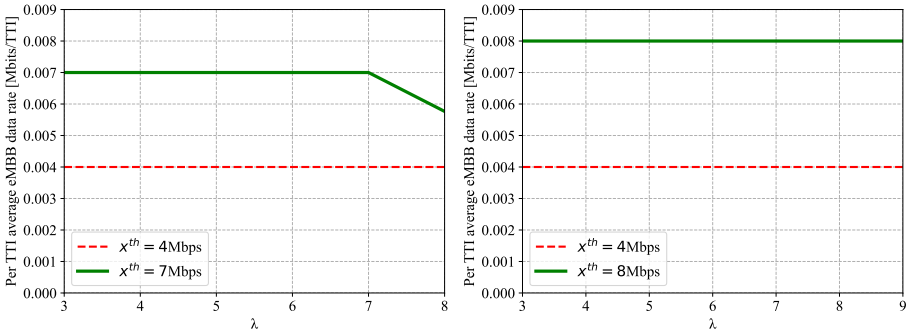
Initially, we assess the delay experienced by uRLLC users under different uRLLC traffic demands. For the first and second scenarios, illustrated in Figures ì2a and 2b the numerology considered is 0 and 1, respectively. In these scenarios, the target eMBB data rate for each UE is fixed to 4 Mbps while the uRLLC traffic demand is varied for all the users. The results underline that,



(a) Numerology μ=0 (b) Numerology μ=1

Figure 2: Average delay for uRLLC UEs [ms] for CERS and SRS, as uRLLC traffic demand (λ) is varied for Numerology μ=0 in (a) and μ=1 in (b).

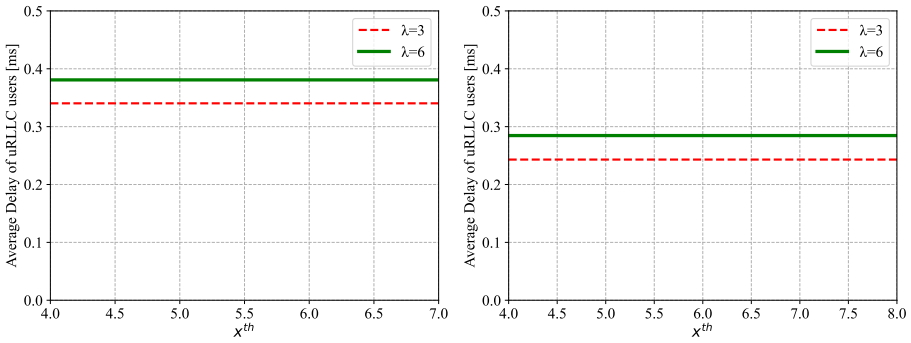
as the uRLLC traffic demand increases, the observed delay for different values of λ always remains below the maximum tolerable delay value (set to 1 ms). However, the uRLLC delays in CERS are higher compared to SRS due to the higher number of RBs allocated under SRS than under CERS. An important result thus follows: at the cost of a slight increase in delay without overshooting the maximum tolerable delay, CERS allows for a considerable reduction in the overall CPU consumption of the RAN compared to its benchmark.



(a) Numerology μ=0 (b) Numerology μ=1

Figure 3: Per-TTI average data rate of eMBB users with two target data rates of 4 and 7 Mbps, respectively, in Fig. 3a, and 4 and 8 Mbps, respectively, in Fig. 3b. The traffic demand of uRLLC users (λ) is varied in both cases.

In the subsequent analysis, we vary the traffic demand of the uRLLC slice while maintaining a constant eMBB traffic demand, hence eMBB target performance. The impact on the achieved data rate of eMBB users is then evaluated for our proposed approach CERS. Fig. 3 illustrates the average observed data



(a) Numerology μ=0 (b) Numerology μ=1

Figure 4: Average delay for uRLLC UEs [ms], as the eMBB target performance (x^{th}) [Mbps] is varied with $\lambda = 3$ and 6, respectively, for Numerology μ=0 in (a), and with $\lambda = 3$ and 6, respectively, for Numerology μ=1 in (b).

rate of the eMBB users with respect to the uRLLC traffic load for two numerologies (namely, 0 and 1). We set x^{th}=4 Mbps and 7 Mbps, respectively, as the two target data rates for each eMBB user in μ=0 (see Fig. 3a).

Subsequently, we analyze the performance of CERS in managing incoming uRLLC load. As observed in Fig. 3a, the eMBB users consistently maintain their target data rate when uRLLC traffic demand (λ) is varied from 3 to 7. However, when the incoming uRLLC traffic demand goes beyond λ=7, the gNB adopts a strategy of puncturing eMBB users to prioritize serving the uRLLC traffic. In this scenario, our proposed approach CERS strives to balance the needs of uRLLC traffic users while minimizing the impact on eMBB users. Consequently, the average achieved data rate of eMBB users is slightly below the target (e.g., achieved eMBB data rate around 5.8 Mbps for λ=8). The achieved data rate may vary based on the considered number of TTIs for calculating the achievable data rate per second. This outcome underscores CERS's capability to either maintain the target data rate or keep it marginally below the target as uRLLC traffic demand rises. In the case of μ=1 (3b), we set a higher target data for each eMBB user equal to 8 Mbps (due to the higher PRB rate in μ=1). Notably, CERS consistently maintains the target data rate even when faced with increasing traffic demand for each uRLLC user. We remark that a lower number of RBs allocated to uRLLC users (as in Fig. 1b) prevents the gNB from puncturing RBs from eMBB users to make room for uRLLC traffic. This strategic allocation ensures that the gNB fulfills the requirements of uRLLC traffic users without compromising the resources allocated to eMBB users.

In our final evaluation, we scrutinize the delay experienced by uRLLC users while varying the traffic demand of the eMBB slice, with the uRLLC slice demand held constant. Fig. 4 illustrates the average delay of uRLLC users as the eMBB traffic demand is varied. The scenarios consider constant uRLLC traffic demand of λ=3 and 6 for Numerology μ=0 in (a) and Numerology μ=1 in

(b), respectively. Remarkably, the observed delay consistently remains below the maximum tolerable delay value of 1 ms, and it remains constant even with higher eMBB rates. It is worth highlighting that, by puncturing the necessary number of RBs from eMBB users, CERS provides the uRLLC users with the necessary RBs to meet their delay requirements while preserving the target eMBB data rate. Also, the delay is consistently lower in higher numerology schemes (Fig. 4b) due to the higher PRB rate in μ=1, and, as expected, the delay increases as λ grows.

6 Conclusion

In this paper, we addressed the cost-efficient resource allocation problem in 5G NR featuring network slicing. We formulated a resource allocation problem that maximizes the slice profit while guaranteeing uRLLC constraints with respect to latency as well as the target data rate of eMBB users. Given the problem inherent complexity, we introduced a strategic approach by decoupling the original problem into two sub-optimization problems, eMBB resource allocation and uRLLC resource allocation. We then used a simple, low-complexity heuristic for the eMBB resource allocation that maximizes the MEAR among eMBB users at time-slot boundaries. Meanwhile, for uRLLC resource allocation at every mini-slot of a time slot, we maximized the slice profit while meeting slice-specific SLAs. Our numerical results demonstrate that our approach achieves cost-efficient resource slicing, and meets the data rate and delay requirements outlined in the SLAs for both eMBB and uRLLC slices.

References

[1] W. Jiang, B. Han, M. A. Habibi, H. D. Schotten, The road towards 6g: A comprehensive survey, IEEE Open Journal of the Communications Society 2 (2021) 334–366. doi:10.1109/OJCOMS.2021.3057679.

[2] S. Zhang, An overview of network slicing for 5g, IEEE Wireless Communications 26 (3) (2019) 111–117.

[3] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, G. S. Paschos, The algorithmic aspects of network slicing, IEEE Communications Magazine 55 (8) (2017) 112–119. doi:10.1109/MCOM.2017.1600939.

[4] P. Popovski, K. F. Trillingsgaard, O. Simeone, G. Durisi, 5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view, IEEE Access 6 (2018) 55765–55779. doi:10.1109/ACCESS.2018.2872781.

[5] 3gpp ts 38.300 v16.8.0, "technical specification group radio access network; nr; nr and ng-ran overall description; stage 2 (release 16),, Tech. rep. (Dec 2021).

[6] 5g america, "new services and applications with 5g ultra-reliable low latency communications," white paper,, Tech. rep. (Nov. 2018).

[7] 3gpp r1-1700374, "downlink multiplexing of embb and urllc transmission,", Tech. rep. (Jan. 2017).

[8] J. Huang, L. Gao, Wireless Network Pricing, Vol. 6, 2013.

[9] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, C. S. Hong, Coexistence mechanism between embb and urllc in 5g wireless networks, IEEE Transactions on Communications 69 (3) (2021) 1736–1749. doi:10.1109/TCOMM.2020.3040307.

[10] A. Anand, G. De Veciana, S. Shakkottai, Joint scheduling of urllc and embb traffic in 5g wireless networks, in: IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, 2018, pp. 1970–1978. doi:10.1109/INFOCOM.2018.8486430.

[11] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, C. S. Hong, embb-urllc resource slicing: A risk-sensitive approach, IEEE Communications Letters 23 (4) (2019) 740–743. doi:10.1109/LCOMM.2019.2900044.

[12] A. Esmaeily, H. V. K. Mendis, T. Mahmoodi, K. Kralevska, Beyond 5g resource slicing with mixed-numerologies for mission critical urllc and embb coexistence, IEEE Open Journal of the Communications Society 4 (2023) 727–747. doi:10.1109/OJCOMS.2023.3254816.

[13] Y. Prathyusha, T.-L. Sheu, Coordinated resource allocations for embb and urllc in 5g communication networks, IEEE Transactions on Vehicular Technology 71 (8) (2022) 8717–8728. doi:10.1109/TVT.2022.3176018.

[14] Y. Huang, S. Li, C. Li, Y. T. Hou, W. Lou, A deep-reinforcement-learning-based approach to dynamic embb/urllc multiplexing in 5g nr, IEEE Internet of Things Journal 7 (7) (2020) 6439–6456. doi:10.1109/JIOT.2020.2978692.

[15] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, C. S. Hong, embb-urllc resource slicing: A risk-sensitive approach, IEEE Communications Letters 23 (4) (2019) 740–743. doi:10.1109/LCOMM.2019.2900044.

[16] P. Yang, X. Xi, T. Q. S. Quek, J. Chen, X. Cao, D. Wu, How should i orchestrate resources of my slices for bursty urllc service provision?, IEEE Transactions on Communications 69 (2) (2021) 1134–1146. doi:10.1109/TCOMM.2020.3038196.

[17] S. Parkvall, E. Dahlman, A. Furuskar, M. Frenne, Nr: The new 5g radio access technology, IEEE Communications Standards Magazine 1 (4) (2017) 24–30. doi:10.1109/MCOMSTD.2017.1700042.

[18] P. Yang, X. Xi, T. Q. S. Quek, J. Chen, X. Cao, D. Wu, How should i orchestrate resources of my slices for bursty urllc service provision?, IEEE Transactions on Communications 69 (2) (2021) 1134–1146.

[19] W. Wu, N. Chen, C. Zhou, M. Li, X. Shen, W. Zhuang, X. Li, Dynamic ran slicing for service-oriented vehicular networks via constrained learning, IEEE Journal on Selected Areas in Communications 39 (7) (2021) 2076–2089. doi:10.1109/JSAC.2020.3041405.

[20] Q. Liu, T. Han, N. Zhang, Y. Wang, Deepslicing: Deep reinforcement learning assisted resource allocation for network slicing, in: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020, pp. 1–6. doi:10.1109/GLOBECOM42002.2020.9322106.

[21] H. Zhang, G. Pan, S. Xu, S. Zhang, Z. Jiang, A hard and soft hybrid slicing framework for service level agreement guarantee via deep reinforcement learning, in: 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), 2022, pp. 1–5. doi:10.1109/VTC2022-Spring54318.2022.9860789.

[22] S. Bakri, P. A. Frangoudis, A. Ksentini, M. Bouaziz, Data-driven ran slicing mechanisms for 5g and beyond, IEEE Transactions on Network and Service Management 18 (4) (2021) 4654–4668. doi:10.1109/TNSM.2021.3098193.

[23] Y. Zhao, X. Chi, L. Qian, Y. Zhu, F. Hou, Resource allocation and slicing puncture in cellular networks with embb and urllc terminals coexistence, IEEE Internet of Things Journal 9 (19) (2022) 18431–18444. doi:10.1109/JIOT.2022.3160647.

[24] S. Pramanik, A. Ksentini, C. F. Chiasserini, Cost-efficient slicing in virtual radio access networks, Computer Communications 209 (2023) 349–358. doi:https://doi.org/10.1016/j.comcom.2023.07.004.

[25] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, C. S. Hong, Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach, IEEE Transactions on Wireless Communications 20 (7) (2021) 4585–4600. doi:10.1109/TWC.2021.3060514.

[26] M. Setayesh, S. Bahrami, V. W. Wong, Resource slicing for embb and urllc services in radio access network using hierarchical deep learning, IEEE Transactions on Wireless Communications 21 (11) (2022) 8950–8966. doi:10.1109/TWC.2022.3171264.

[27] K. Boutiba, M. Bagaa, A. Ksentini, Optimal radio resource management in 5g nr featuring network slicing, Computer Networks, Vol. 234, October 2023 (2023).

[28] H. Yang, K. Zheng, K. Zhang, J. Mei, Y. Qian, Ultra-reliable and low-latency communications for connected vehicles: Challenges and solutions, IEEE Network 34 (3) (2020) 92–100.

[29] S. Pramanik, A. Ksentini, F. Chiasserini, C. Characterizing the computational and memory requirements of virtual rans, in: 2022 17th Wireless On-Demand Network Systems and Services Conference (WONS), 2022, pp. 1–8.

[30] S. Pramanik, A. Ksentini, C. F. Chiasserini, Cost-efficient slicing in virtual radio access networks, Computer Communications 209 (2023) 349–358. doi:https://doi.org/10.1016/j.comcom.2023.07.004.

[31] `https://en.wikipedia.org/wiki/Divide-and-conquer_algorithm`.

[32] https://www.gurobi.com/.