User-Based Quality of Service Aware Multi-Cell Radio Access Network Slicing

Arled Papa, Alba Jano, Serkut Ayvaşık, Onur Ayan, H. Murat Gürsu, Wolfgang Kellerer
Chair of Communication Networks, Technical University of Munich
{arled.papa, alba.jano, serkut.ayvasik, onur.ayan, murat.guersu, wolfgang.kellerer}@tum.de

***Abstract*—5G radio access network (RAN) slicing envisions a solution to flexibly deploy heterogeneous services as slices sharing the same infrastructure. However, this level of flexibility renders slice isolation challenging, mainly due to the stochastic nature of wireless resources. In the state-of-the-art, RAN slicing algorithm's efficiency with respect to slice isolation is related to the ability of meeting individual slice requirements. However, mostly an aggregated slice performance guarantee is considered instead of per user guarantees. Hence, state-of-the-art approaches might not always provide the satisfaction of all users within a slice. Indeed, our results demonstrate that if user requirements within a slice are not included in the RAN slicing algorithm, the per user quality-of-service (QoS) may not be fulfilled. In this paper, we investigate the definition of slice isolation as the ability to satisfy individual users' throughput within slices, in a frequency selective, multi-cell wireless scenario with focus on maximizing slices' throughput. Our problem is tackled with a Lyapunov optimization approach, which proves to always achieve slice isolation. Our results show that our solution does not only achieve 100% user QoS guarantees compared to 50% achieved in the state-of-the-art, but also doubles the throughput with increasing number of BSs.**

***Index Terms*—RAN Slicing, Slice isolation, 5G, SD-RAN.**

I. Introduction

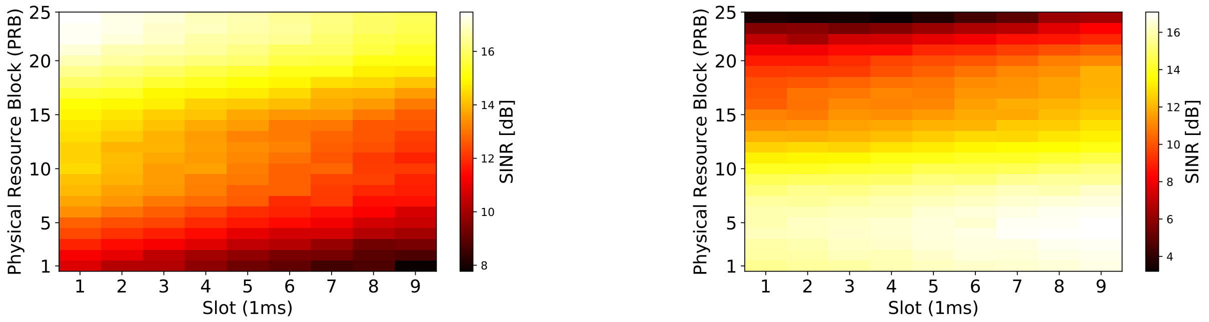
Next generation 5G/6G networks are facing increased traffic demands and a rise of heterogeneous applications such as tele-operations, online gaming, aircraft entertainment, in-train communications and internet of things (IoT) with thousands of connected devices [1]. Existing one-size-fits-all network architectures cannot cater for this level of heterogeneity. To this end, 5G/6G networks are witnessing a new era of radio access networks (RANs). The concepts of programmability and softwarization are being introduced by 3GPP standardization [2] to offer new levels of flexibility. In that regard, the conventional RAN architecture has been split into control and data plane through means of software-defined networking (SDN), paving the way towards SD-RAN architectures. The SD-RAN architecture envisions a centralized control plane in what called SD-RAN controllers and the data plane remains in the base stations (BSs). This centralization enables the SD-RAN controllers to have a better view of the underlying BSs and achieve higher performance through careful coordination.

While SDN can ease the resource management and orchestration, still cannot accommodate standalone the heterogeneity of emerging applications. To address the above issue, the concept of RAN slicing has emerged. RAN slicing enables the co-existence of multiple virtual mobile network operators (VMNOs) or third parties such as automotive, factories, aircrafts, sharing the same physical infrastructure, where each slice should remain unaffected by other slices, preserving isolation [3].

While RAN slicing offers more efficient spectrum allocation (i.e., physical resource blocks (PRBs), cost reduction and multiplexing gain [4], [5], [6], it renders slice isolation challenging, especially in a wireless environment due to the stochastic nature and scarce wireless resources. The problem of slice isolation has been mainly tackled for a single cell scenario [7], [8]. Nonetheless, in reality a cellular network consists of multi-cells. Considering, a multi-cell scenario, the slice isolation preservation becomes even more challenging due to the limited frequency bands [9], resulting in interference introduced by slices being deployed in adjacent BSs. There already exist interference management techniques in 5G [10], [11], [12] with respect to multi-cell scenarios, but none of them considers the problem of RAN slicing, which is more challenging especially due to a lack of information shared among different network slices operated individually by VMNOs [13].

When considering slice isolation, in general there are two questions that need to be answered: **i) What is the right metric for isolation? and ii) What should be the isolation granularity: slice based or user based?** Regarding question (i) the literature is divided into 2 main categories, namely radio resource-based isolation [13], [14] i.e., if a slice is provided a specific amount of resources, and performance-based isolation [7], [8], [15] i.e., if a slice fulfills a minimum quality-of-service (QoS) requirement. While the former would be sufficient for flat-fading wireless channels, in reality the latter is beneficial to consider, as it takes into account QoS requirements and therefore can benefit from diversity gains in case of frequency-selective wireless channels. Considering question (ii), the state-of-the-art is also divided into two groups. For instance, there are works that define the radio resource-based and performance-based isolation mainly on a slice level [16], [17], [18], whereas there are other works based on user level [7], [8], [19]. We stress that while the first approach provides QoS for each user when users depict homogeneous characteristics, it cannot provide user QoS if users experience different channel characteristics. We conclude that in order to guarantee user QoS, an isolation at user level is mandatory.

To illustrate the above points let us consider Fig. 1. In the figure, the wireless channel quality differences are depicted in



(a) t=0 ms

(b) t=240 ms

Figure 1: Received SNR of one user for one LTE Frame in 5 MHz Bandwidth at different time instances in low doppler frequency EPA channel.

terms of received signal to noise ratio (SNR) for one user at different time instances in a low mobility i.e., 5 Hz doppler frequency extended pedestrian a (EPA) channel [20] in 5 MHz bandwidth corresponding to 25 PRBs. The SNR is displayed as a colorbar where x-axis represents time in slots and y-axis corresponds to frequency in PRB notation. Considering the relation between SNR and reliability, it is apparent that due to the frequency selective fading, the allocation of PRBs between 1 and 5 result in a different throughput than an allocation of PRBs between 20 and 25, although for one user, 20% of the resources would have been assigned in both cases. Furthermore, the SNR and frequency-time grid relation changes over time due to user mobility. Fig. 1a and Fig. 1b illustrate the same wireless channel model realization, but with 240 ms time difference. Thus, both frequency and time selectivity of the channel should be considered in the resource allocation optimization. In fact, in [21] it is shown that wireless channel dependent scheduling can increase the achieved throughput up to 80% compared to static scheduling. Hence, approaches that tend to neglect the wireless channel effects diminish the effectiveness of the RAN slicing optimization.

Given all the above mentioned issues, in this paper we identify the definition of slice isolation on a per user basis as necessary to provide actual guarantees in the network, while considering a multi-cell scenario. In more detail the contributions of this paper are as follows:

1. We propose a RAN slicing isolation definition based on performance guarantees on a per user basis to cater for frequency selective wireless channels and time variations.
2. We develop and provide a mathematical formulation of RAN slicing and isolation in a multi-cell scenario with slices being deployed in adjacent interfering BSs with focus on throughput maximization, while satisfying user constraints. Due to the complexity of the problem, we propose an approximation solution based on Lyapunov optimization.
3. We demonstrate the effectiveness of our approach with respect to users and BSs as well as the convergence time of the algorithm. To verify our results, we consider an aircraft in-cabin channel model representing a 5G multi-cell use-case with high user density.
4. We compare our algorithm with existing state-of-the-art solutions with respect to network slice isolation and throughput maximization, as well as CPU, memory utilization and communication overhead.

The structure of the paper is organized as follows: We provide a comprehensive state-of-the-art analysis on RAN slicing both on single and multi-cell scenarios in Section II. Section III introduces the system model and details the optimization problem, whereas the approximation solution based on the Lyapunov optimization is described in Section IV. We demonstrate the results of our solution and comparisons with other approaches in Section V. Finally, we discuss our work and conclude our paper in Section VI.

II. Related work

The problem of RAN slicing has experienced large attention in the last couple of years with vast ongoing research. Valuable conceptual works regarding RAN slicing can be found in [4], [5], [6], [22], [23], [24]. From the practical perspective, the new concept of decoupling the control and data plane of conventional RANs facilitates the development of prototype SD-RAN platforms [25], [26], [27], which provide a valuable asset for academia to deploy and test algorithms in realistic scenarios. Nonetheless, only a limited number of users (UEs) and base stations (BSs) can be tested due to expensive radio equipment. From the mathematical perspective the RAN slicing problem has been addressed mainly for a single BS scenario ranging from optimizations [7], [8], [14], [17], [18], [19], [28], [29] to genetic algorithms [30], physical layer perspective [31] as well as the business model aspect [32], [33]. Furthermore, deep reinforcement learning has been utilized in [34], which adapts to dynamic traffic demands and preserves long term QoS requirements. From the industrial applications' perspective authors in [35] consider a network slicing use case under deterministic traffic assumptions, demonstrating higher throughput compared to alternative solutions. While the aforementioned works are valuable and proven efficient for a single BS scenario, they cannot be used for a multi-BS scenario, where the interference among adjacent BSs has to be considered carefully when allocating resources to slices.

Similar to our work, there are papers considering RAN slicing in a multi-cell scenario. Authors in [36] provide a QoS preservation scheduling approach for heterogeneous traffic requirements and prove to be efficient. Nonetheless they do not specifically provide evaluations on a per user level QoS. In [16] a RAN slicing solution is provided for multi-cells while maximizing the spectral efficiency. However, the granularity remains on a slice level and no details about the user throughput are provided. Considering a frequency selective channel not all users are treated equally. Work [37] presents different possibilities for the RAN slicing over multiple BSs and demonstrates a qualitative representation of the isolation effect of the approaches. However, the QoS is portrayed on a slice level and not on a user level. Works [15], [38] consider a multi-cell scenario providing an algorithm for dynamic resource allocation among slices. Moreover, they compare their approach with two baselines solutions considering static slicing. Yet, they do not address directly the network slice isolation problem. Alternatively, authors in [39] combine the user admission with the resource scheduling problem to achieve user QoS requirements. Nonetheless, their optimization problem focuses on minimizing resource consumption while satisfying QoS requirements, rather than maximizing the overall system throughput. The closest work to our approach is provided in [13]. In [13] a method to enforce the network slicing policies of virtual mobile network operators (VMNOs) over multi-BSs is suggested, while providing a percentage of resources as the slice isolation criteria. While this might be sufficient for flat-fading selective wireless channels, where all the resources are the same for all users, it might not be optimal in a frequency selective and time-variant case. In such scenarios selecting the best channel per user within a slice increases the network performance. In contrary, if a bad channel is assigned to a network slice, even if the resource percentage is achieved no user QoS can be guaranteed. Motivated, by such scenarios, in this work we propose a scheme which is more granular by offering a minimum achievable rate per user within a slice and concluding about the slice isolation effect.

III. System model and Problem Formulation

We consider a downlink scenario of a cellular network consisting of a set $\mathcal{B}$ of B base stations (BSs). Following the realistic assumption of limited spectrum bands for commercial use [9], in our scenario adjacent BSs are interfering with each other. The radio access network (RAN) belongs to a single network provider (NP), but can be leased to other virtual mobile network operators (VMNOs) or third parties (i.e., aircraft, automotive, healthcare) referred to as slice owners or tenants, in terms of network slices. We assume that in total in our system we have a set $\mathcal{S}$ of S network slices. In turn, network slices can be deployed in multiple BSs. The RAN serves a total number of N users (UEs) from a set $\mathcal{N}$. We define δ_i^s as a binary variable being 1 if user i belongs to slice s and 0 otherwise. This variable is already predefined by the slice owner. Each user in the network can be attached to only one network slice s i.e., $\sum_s \delta_i^s = 1 \quad \forall i \in \mathcal{N}$ and it can only be served by one BS b at a time. We consider a time slotted system of T slots in total, where each slot corresponds to the transmission time interval (TTI), which is 1 ms in LTE. The spectrum is divided into a set $\mathcal{R}$ of R physical resource blocks (PRBs) [40].

Let $h_{i,j}^b(t)$ represent the channel gain of user $i \in \mathcal{N}$, on PRB $j \in \mathcal{R}$ for base station $b \in \mathcal{B}$ in slot $t \in T$. For the sake of representing Rayleigh block fading, we assume that the channel gain remains constant for the coherence time of the channel τ and changes every multiple of τ for all the users in the system. In this paper, τ corresponds to 20 ms as reported in [41], representing dynamic channel conditions. Let variable $w_{i,j}^b(t)$ be a decision binary variable taking the value 1 if user $i \in \mathcal{N}$ has been assigned PRB $j \in \mathcal{R}$ for base station $b \in \mathcal{B}$ in slot $t \in T$, or 0 otherwise. We then define $w_{i,j}^{b,s}(t) = w_{i,j}^b(t) \cdot \delta_i^s$ as the outcome of the resource allocation per slice. We can write the rate that each user achieves in slot t as follows:

$$r_i(t) = \sum_{b=1}^{B} \sum_{s=1}^{S} \sum_{j=1}^{R} w_{i,j}^{b,s}(t) \cdot r_{i,j}^b(t). \tag{1}$$

Let variable $p_j^b(t)$ denote the power allocation for the transmission on PRB $j \in \mathcal{R}$ of base station $b \in \mathcal{B}$ in slot $t \in T$. Assuming a uniform distribution of the power along all PRBs in a BS b and denoting the variable N_0 as the thermal noise we can write:

$$r_{i,j}^b(t) = W \log_2 \left(1 + \frac{h_{i,j}^b(t) \cdot p_j^b(t)}{I + N_0} \right). \tag{2}$$

From equation (2), W is the bandwidth of each PRB i.e., 180 KHz in LTE, whereas the interference experienced by each user denoted by I is calculated as:

$$I = \sum_{\beta \in \mathcal{B} \setminus \{b\}} h_{i,j}^{\beta}(t) \cdot p_j^{\beta}(t), \tag{3}$$

and it is composed by the rate of each interfering BS to user i served by BS b.

We present the architectural concept of our solution which is based on the principles of SD-RAN as elaborated in [25], [26], [27] with the help of Fig. 2. Initially, we detail our envisioned SD-RAN solution. Then, we describe the interaction of VMNOs or third parties with the SD-RAN controller utilizing RAN slicing requests and finally we present the algorithm to perform RAN slicing.

A. SDN-enabled RAN slicing

This subsection envisions our SD-RAN solution for RAN slicing. Our concept follows a hierarchical scheduling approach similar to [19], [23], where a master controller is based on the SD-RAN controller, whereas each slice is managed by a scheduling entity referred to as **slice manager** as portrayed in Fig. 2.

The SD-RAN controller is the heart of the SD-RAN platform managing the interaction with the VMNOs or third parties and the slice managers itself. It receives requests from the slices/tenants from the north bound application programmable

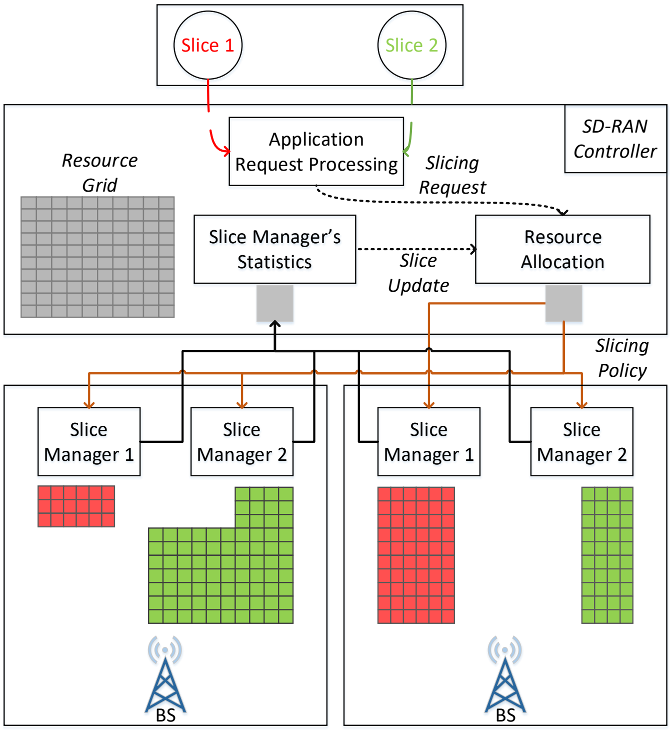


Figure 2: Multi-cell SD-RAN enabled RAN slicing architecture. The SD-RAN controller is the heart of the architecture. For every slice, a slice manager is created that communicates with the SD-RAN controller and provides efficient resource allocation for network slices.

interface (API). These requests are utilized as inputs for the algorithm described in the following and influence the resource allocation. The output of the resource allocation entails which resources should be given to which slice and to which BS as illustrated in Fig. 2.

Within the SD-RAN controller we propose three main functionalities, which render an efficient RAN slicing. Initially, an **application request processor** gathers and distributes the RAN slicing requests from the tenants to the **resource allocator** block. The resource allocator is in charge of dynamically re-assigning resources to slice managers within each BS. Finally, to alleviate the RAN slicing process, a **slice manager statistics** block contains all the information with respect to users within a slice and their respective channel conditions. The SD-RAN controller utilizes these statistics to perform a better allocation. The frequency of statistics updates depends on the channel coherence time τ. It is important to contain up-to-date statistics every time the channel changes in order to perform the best allocation. In this work, updates are performed every 20 ms as measured in [41], demonstrating dynamic channel variations. The SD-RAN controller dynamically re-assigns resources to each slice manager, whereas the latter is responsible for allocating them to users within that slice. The duty of the SD-RAN controller is to cooperate effectively with the slice managers in order to optimize the network. The full details of the optimization process will be given in Section IV.

B. RAN Slicing Request

Each slice s communicates to the SD-RAN controller its requests through the northbound API. As explained earlier in this paper these requests are presented either in terms of minimum radio resource requirements or minimum QoS requirements. In turn, this can be expressed on a per user level, or on a per slice level. Let $\rho_s \in [0,1]$ express the percentage of resources requested by slice s such that $\sum_{s\in\mathcal{S}} \rho_s = 1$. Each slice has an average traffic arrival λ_s in every slot t, whereas the system admits traffic portion wise with an average rate α_s per slot. Moreover, let $K_s^t = \sum_{b=1}^{B}\sum_{i=1}^{N}\sum_{j=1}^{R} w_{i,j}^{b,s}(t)$ be the total amount of PRBs allocated to slice s in slot $t \in T$. Finally, let R be the total amount of PRBs per BS and η_s^b a binary variable that denotes whether a slice is allowed to be served by BS b or not.

Definition 1. *(Aggregate radio resource requirement). Given a minimum radio resource requirement per slice ρ_s, network slices are considered isolated if:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} K_s^t \geq \rho_s \cdot R \cdot \sum_{b=1}^{B}\eta_s^b. \tag{4}$$

Definition 2. *(QoS requirement per slice). Given a minimum QoS rate per slice $\underline{C_s}$, network slices are considered isolated if:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{b=1}^{B}\sum_{i=1}^{N}\sum_{j=1}^{R} w_{i,j}^{b,s}(t)\cdot r_{i,j}^{b}(t) \geq \underline{C_s} \quad \forall s\in\mathcal{S} \tag{5}$$

Definition 3. *(QoS requirement per user). Given a minimum QoS rate per user $\underline{C_s^i}$, network slices are considered isolated if:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{b=1}^{B}\sum_{j=1}^{R} w_{i,j}^{b,s}(t)\cdot r_{i,j}^{b}(t) \geq \underline{C_s^i} \quad \forall i\in\mathcal{N}, \forall s\in\mathcal{S} \tag{6}$$

C. RAN Slicing Problem Formulation

To maximize the number of users served in a slice, in our work, the RAN slicing algorithm aims at maximizing the overall slice throughput. As aforementioned, we follow Definition 3 for assuring slice isolation, where we guarantee a minimum throughput requirement for the slice's users. The optimization problem is defined as follows:

$$\mathcal{P}_0 : \max_{w^b_{i,j}(t)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{b=1}^{B}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{R} w^{b,s}_{i,j}(t)\cdot r^b_{i,j}(t) \tag{7}$$

$$\text{s.t.} \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\alpha_s(t) \leq \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} r_s(t) \quad \forall s\in\mathcal{S} \tag{8}$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{b=1}^{B}\sum_{j=1}^{R} w^{b,s}_{i,j}(t)\cdot r^b_{i,j}(t) \geq \underline{C^i_s}$$

$$\forall i\in\mathcal{N}, \forall s\in\mathcal{S} \tag{9}$$

$$\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{R} w^b_{i,j}(t) \leq R \quad \forall t\in T, \forall b\in\mathcal{B} \tag{10}$$

$$\sum_{b=1}^{B}\eta^b_s \leq B_s \quad \forall s\in\mathcal{S} \tag{11}$$

$$\sum_{s=1}^{S}\sum_{i=1}^{N_s} w^b_{i,j}(t) \leq 1 \quad \forall t\in T, \forall b\in\mathcal{B}, \forall j\in\mathcal{R} \tag{12}$$

$$0\leq \alpha_s(t) \leq \lambda_s(t) \quad \forall t\in T, \forall s\in\mathcal{S} \tag{13}$$

$$w^b_{i,j}(t)\in\{0,1\} \tag{14}$$

$$\eta^b_s\in\{0,1\} \tag{15}$$

Constraint (8) assures that the average rate of each slice is greater than the average admitted traffic for each slice. In order to preserve equality among users within a slice and satisfy each user's requirements we introduce constraint (9), that guarantees that the average rate of each user within a slice is greater than a minimum threshold assigned by the slice owner i.e., $\underline{C^i_s}$. Each BS is limited to an amount of PRBs of size R. That means that for all users of slices served by BS b no more than R resources can be given. This is assured by constraint (10). Given that slices are distributed over multi-cells, we define the variable η^b_s as a binary variable being 1 if a slice is allowed to be placed on BS b or 0 otherwise. We can limit the number of BSs that each slice can be assigned to i.e., B_s, by using constraint (11). The orthogonality constraint for each PRB is finally guaranteed by constraint (12). Constraint (13) suggests that the total admitted traffic cannot exceed the average arrival rate of each slice in order to operate in a stable region.

The main challenge of optimally solving $\mathcal{P}_0$ is due to the stochastic nature of the problem. That said, the lack of knowledge with respect to wireless channel and time variations as well as user traffic requests over time constitutes the problem difficult to solve with traditional linear programming methods. Even with knowledge about the future, with increasing T the problem becomes hard to solve. The aforementioned challenges indicate that in order to solve $\mathcal{P}_0$, an online method that can provide suboptimal solution based only on per slot information is mandatory. Lyapunov optimization has been suggested and proven efficient for the aforementioned problems [42], thus we utilize it in our work to solve $\mathcal{P}_0$.

IV. Approximation solution with Lyapunov optimization

In this section, we detail the Lyapunov optimization approach selected for solving the RAN slicing problem $\mathcal{P}_0$. As slice isolation is one of the main challenges in RAN slicing, a technique to obtain the isolation constraint as part of the objective function is needed. Following the Lyapunov optimization approach, the constraints are transformed into virtual queues and they become part of the objective function. The virtual queues of the problem evolve over time as follows:

$$U_s(t+1) = max\{U_s(t)+\alpha_s(t)-r_s(t), 0\} \quad \forall s\in\mathcal{S} \tag{16}$$

$$L_i(t+1) = max\{L_i(t)+\underline{C^i_s}-\sum_{b=1}^{B}\sum_{j=1}^{R} w^{b,s}_{i,j}(t)\cdot r^b_{i,j}(t), 0\}$$

$$\forall i\in\mathcal{N}, \forall s\in\mathcal{S} \tag{17}$$

The virtual queue $U_s(t)$ is a slice based queue that indicates the physical backlog queue of the system, which evolves according to the admitted traffic and the served traffic on every slot $t \in T$. Alternatively, the virtual queue $L_i(t)$ represents the user queue that is related to the user minimum throughput requirement stated by constraint (9) in $\mathcal{P}_0$. Using the virtual queues we define an overall system queue state $\Theta(t) = \{\Theta_s(t)\}$ where $\Theta^t_s = \{U_s(t), L_i(t)\}$, indicating the state of slice s. The quadratic Lyapunov function is then defined as follows:

$$L(\Theta(t)) = \frac{1}{2}\sum_{s\in S}(U_s(t))^2 + \frac{1}{2}\sum_{s\in S}\sum_{i\in N}(L_i(t))^2 \quad \forall t\in T \tag{18}$$

The Lyapunov function is a scalar metric to measure the *queue congestion* state. A small value of $L(\Theta^t)$ implies that the stability of the queues holds and therefore the constraints are satisfied. In the contrary, if the algorithm cannot satisfy the constraints then a large value of $L(\Theta(t))$ is observed and as a result we conclude that the system is not stable. In any case, all constraints are satisfied if, and only if, the infinite time-horizon limit of $L(\Theta(t))$ is bounded, i.e., $\lim_{T\to\infty} 1/T\sum_{t=0}^{T-1} L(\Theta(t)) < \infty$. According to [42], to enforce stability the one-slot Lyapunov drift of the Lyapunov function is checked in every slot t as follows:

$$\Delta L(\Theta(t)) = \mathbb{E}\left\{L(\Theta(t+1)) - L(\Theta(t))|\Theta(t)\right\} \tag{19}$$

Based on the Lyapunov technique, the objective is to minimize an infinite bound on the Lyapunov drift in each time slot, where the drift-plus-penalty is expressed as:

$$\Delta L(\Theta(t)) - V\mathbb{E}\{\sum_{b=1}^{B}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{R} w^{b,s}_{i,j}(t)\cdot r^b_{i,j}(t)|\Theta(t)\} \tag{20}$$

$V \geq 0$ is the Lyapunov design parameter, which controls the emphasis given to the maximization problem compared to the queue stability. That means, a larger V pushes the algorithm towards optimality, but increases the time needed for queues to converge. In our case, this is directly translated to the ability of users receiving the minimum required throughput (i.e., preserve slice isolation). Referring to [42, Lemma 4.6], the objective is to minimize a bound on the drift-plus-penalty expression that satisfies the following constraint:

$$\begin{aligned}&\Delta L(\Theta(t)) - V\mathbb{E}\{\sum_{b=1}^{B}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{R} w_{i,j}^{b,s}(t)\cdot r_{i,j}^{b}(t)|\theta(t)\} \leq \beta \\ &- V\mathbb{E}\{\sum_{b=1}^{B}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{R} w_{i,j}^{b,s}(t)\cdot r_{i,j}^{b}(t)|\theta(t)\} + \\ &\mathbb{E}\{\sum_{s\in S}(\alpha_s(t) - r_s(t))U_s(t)\} + \\ &\mathbb{E}\{\sum_{i\in N}(\underline{C_s^i} - \sum_{b=1}^{B}\sum_{j=1}^{R} w_{i,j}^{b,s}(t)\cdot r_{i,j}^{b}(t))L_i(t)\}\end{aligned} \tag{21}$$

Our proposed RAN slicing approach maximizes the right hand side of (21), subject to the constraints (10) - (15).

Given that the right hand side of (21) is an integer linear problem we cannot prove convexity unless the decision variable $w_{i,j}^{b}(t)$ is relaxed to continuous such that $w_{i,j}^{b}(t) \in [0,1]$. Due to the fact that the expression on the right-hand side of (21) and the constraints (10) - (13) are convex with respect to the relaxed continuous decision variable $w_{i,j}^{b}(t)$, we derive the expression with respect to the resource allocation $w_{i,j}^{b}(t)$ to find its maximum:

$$\frac{\partial L}{\partial w_{i,j}^{b}(t)} = (-V)r_{i,j}^{b}(t) - r_{i,j}^{b}(t)(U_s(t) + L_i(t)). \tag{22}$$

The derivative presents the dynamic resource allocation algorithm, where for each PRB, the user with a minimum value of $(-V)r_{i,j}^{b}(t) - r_{i,j}^{b}(t)(U_s(t) + L_i(t))$ is selected to be served. Finally, the slice manager updates the virtual queues and the same procedure occurs in the next slot.

V. Results

In this section we demonstrate the main findings of our work. Given that slice isolation is the main focus of our paper, initially we demonstrate why a radio resource-based solution is not sufficient for frequency selective wireless channels and highlight the importance of considering a QoS based isolation definition for network slices.

As earlier mentioned, slice isolation preservation becomes even more challenging when a multi-cell scenario is considered. Moreover, when the network dimensions increase not only in terms of base stations (BSs), but also users (UEs) the ability to preserve isolation is harder. In that regard, in this section we show the performance of the Lyapunov optimization while increasing network dimensions, specifically with respect to UEs and BSs. Given that the Lyapunov optimization is an approximation technique, there is a trade-off between optimality and convergence time of the algorithm. In the remainder of this paper, we will refer to convergence time as the time instance within the simulation, where all the Lyapunov queues are stabilized, indicating a satisfaction of the QoS requirements. As detailed in Section IV the Lyapunov parameter V indicates this trade-off. We additionally provide results with respect to V, which gives a hint about the fine-tuning of this parameter for a specific system setup. Finally, we compare the Lyapunov approach with alternative state-of-the-art solutions in terms of throughput maximization and ability to satisfy user requirements in order to demonstrate the effectiveness of our approach. Given the importance of small cells for 5G networks [43], [44], [45] and the interference including BS density, for our evaluations we consider an in-aircraft channel model as a representative 5G small cell use-case based on realistic measurements performed within a cabin [46]. Although results are presented for an in-aircraft use-case, we stress that our approach is applicable and beneficial for any multi-cell scenario with time and frequency channel variations.

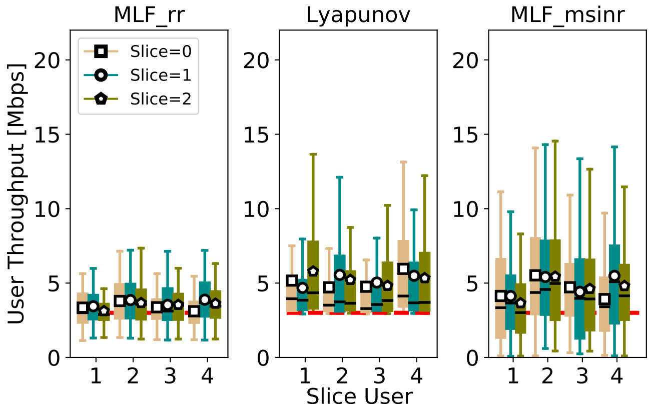


Figure 3: Isolation comparison for MLF_rr, MLF_msinr and Lyapunov approach for 3 Mbps minimum user requirement in a 2 BS scenario. While Lyapunov can effectively satisfy all UEs, MLF_rr and MLF_msinr cannot satisfy all UEs for all runs.

Our simulation environment is Matlab-based, running on a Dell server Intel(R) Xeon(R) E5-2650 CPU, containing 12 physical cores at frequency 2.20 GHz and 64 GB of RAM. Table I represents the main parameters of our simulations. Users are distributed according to the seat plan of Boeing B737-400 [47], where 156 passengers are placed in 26 rows of 6 seats each. For every user within the simulation, channel characteristics are generated according to the presented channel model in Table I. These characteristics change every 20 ms, representing dynamic channel conditions. The simulation contains 50 repetitions, each 5000 slots, where the Lyapunov optimization is utilized to allocate resources to users.

A. Isolation Comparison

The initial results of our work aim at identifying the importance of considering a QoS based slice criteria compared to a radio resource-based one. For that purpose, we compare our algorithm with one of the most relevant approaches in the state-of-the-art with respect to radio resource-based solutions [13]. In [13] the authors introduce the concept of linked PRBs, meaning that identical PRBs should be allocated to the same slice within multi-cells in order to avoid interference by use of sophisticated power management techniques. In that

Table I: Simulation parameters

Parameters	Value
System bandwidth W per BS	5 MHz
Carrier Frequency	1.8 GHz
BS antenna horizontal/vertical	70°/10° beam width
BS antenna down-tilt	15°
Number of PRBs per BS	25
Transmission power per BS	-15 dBm
Shadowing	Gaussian zero-mean with 4.8 dB standard deviation (σ_L)
Free space loss at d_r	37.5 dB
Reference d_r	1 m
Pathloss exponent (n)	2.6
Multipath	Rice distribution with -1.4 dB mean and K-factor 8.1 dB
Path loss in dB at distance d	$PL = F_{d_r} + 10n\log(\frac{d}{d_r}) + X(0,\sigma_L)$
Number of Slices	3
Number of BSs	$[2,3,4,5]$
Number of users per slice	$[4,5,6,7]$
Number of slots	5000

regard, they propose a heuristic approach, namely most linked first (MLF), which starts resource allocation with those slices that have the most linked resources. While, the inter-slice allocation problem is solved with MLF, intra-slice scheduling i.e., how the resources are distributed to users within a slice in work [13] is not explicitly mentioned. For a fair comparison, we define two approaches, namely MLF round robin (MLF_rr) and MLF maximum SINR (MLF_msinr), where the second algorithm defines the way how resources are distributed within a slice.

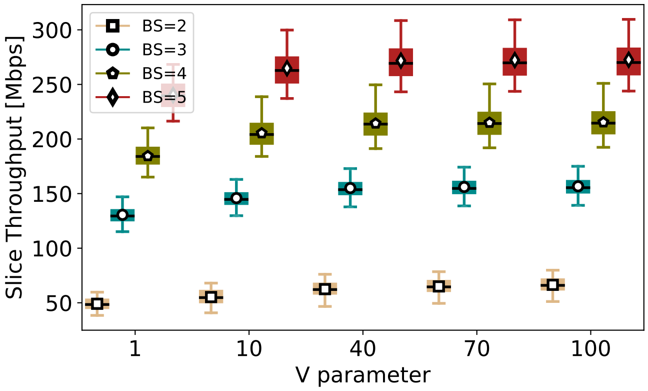


Figure 4: Aggregated slice throughput for Lyapunov optimization with increasing number of BSs for 12 UEs and 3 slices for various V shape parameters. The overall slice throughput increases with increasing V and number of BSs.

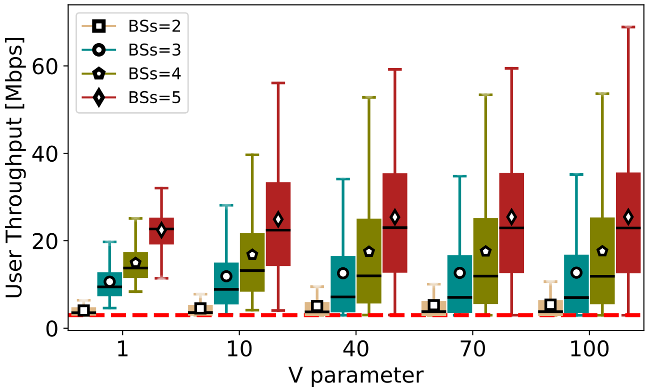


Figure 5: Individual UE throughput for Lyapunov optimization with increasing number of BSs for 12 UEs and 3 slices for various V shape parameters with a minimum requirement of 3 Mbps per UE. Lyapunov optimization can effectively preserve a minimum throughput for all UEs.

Fig. 3 demonstrates the difference between the Lyapunov optimization and RAN slicing enforcement MLF [13] in a frequency selective wireless environment. The x-axis represents a user within a slice and the y-axis represents the achieved throughput for each user. In total there are 12 UEs distributed equally over 3 slices served by 2 BSs. For the MLF approach a minimum requirement per slice is assumed. Given that all slice requirements are identical, the number of resources is equally distributed among slices. On the other hand, for the Lyapunov approach representing a per user-based scheme, a minimum requirement i.e., 3 Mbps per user is selected based on a challenging feasible scenario generated from simulation results. The simulations are performed for 50 runs, where each run consists of 5000 time slots with time and frequency variations every 20 ms. The boxplots for all runs are shown in Fig. 3. The results illustrate that for MLF_rr and MLF_msinr none of the UEs achieves the minimum QoS requirements in all runs. Although a minimum requirement in terms of resources is provided to slices, omitting user channel specific characteristics from the resource allocation leads to throughput degradation, which can be avoided by assigning carefully the most suitable channel per user. Alternatively, the Lyapunov optimization can serve all UEs while satisfying their minimum QoS requirements for all runs. We can therefore conclude that in a frequency selective wireless environment a radio resource-based solution cannot guarantee a minimum QoS requirement for all UEs, thus a QoS based approach is mandatory for the RAN slicing problem to provide isolation.

B. *Impact of number of BSs*

Further investigations are conducted for the proposed Lyapunov optimization with respect to the number of BSs in the network. In principle, when the number of BSs increases in the network and given the limited amount of frequency bands, the probability of BSs interfering with each other also increases. The interference becomes even more noticeable especially considering a small cell scenario like inside an aircraft cabin, where BSs are placed closer to each other, thus producing higher interference. To that end, the preservation of isolation becomes more challenging, thus a careful RAN slicing algorithm should be considered.

In order to demonstrate the effectiveness of our approach we illustrate the outcome of adding BSs in the network with respect to network slices' throughput, while keeping the number of UEs fixed to 12 over all slices. The results are depicted

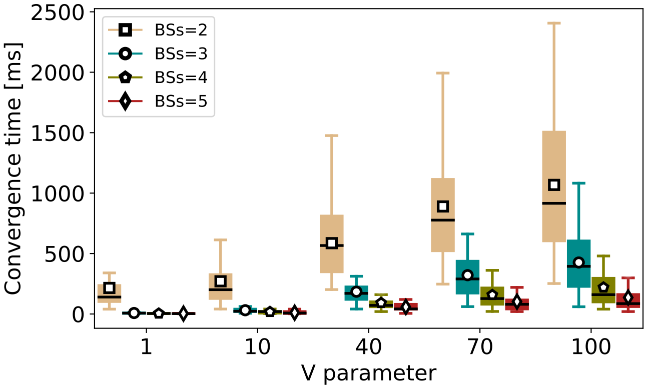


Figure 6: Converge time for Lyapunov optimization with increasing number of BSs for 12 UEs and 3 slices for various V shape parameters. The convergence time of the algorithm increases with increasing V, but decreases with number of BSs due to higher achievable throughput.

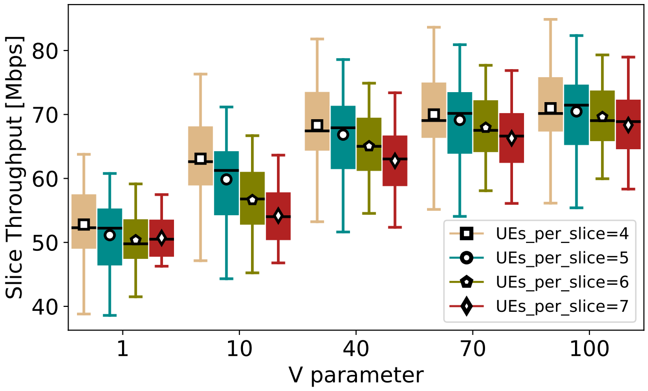


Figure 7: Aggregated slice throughput for Lyapunov optimization with increasing number of UEs for 2 BSs and 3 slices for various V shape parameters. The overall slice throughput increases with increasing V and decreases with number of UEs as it is harder to fulfill the user requirements.

in Fig. 4 and Fig. 5 respectively. The x-axis represents the V shape parameter of the Lyapunov optimization. This parameter depicts the trade-off between optimality i.e., increasing slices' throughput and queue stability, which in our case directly translates to the ability to satisfy the constraints i.e., preserve the minimum throughput requirement per each user. The y-axis shows the total slice throughput in Mbps for Fig. 4 and the individual UE throughput in Mbps for Fig. 5. The simulations have been conducted for 50 runs and the boxplots are drawn to show the distribution of the results. Given that all slices are identical we combine all the UEs of slices into single boxplots to illustrate the results of 50 runs. For the simulations the minimum requirement per each UE within the slice has been fixed to 3 Mbps, whereas the slice arrival λ_s is also 3 Mbps on average. As it can be observed by Fig. 4, the overall slice throughput is increasing with increasing V parameter as expected. Moreover, by increasing the number of BSs, the slice throughput increases. The individual UE throughput varies on average between 4 and 6 Mbps for 2 BSs ranging V from 1 to 100. Alternatively, for 5 BSs the individual UE throughput varies between 22 and 26 Mbps on average ranging V from 1 to 100.

As mentioned above, while interference increases for increasing number of BSs, also the distance to the UEs of each slice decreases. In that regard, a careful allocation of PRBs demonstrates that the interference problem can be omitted and the diversity gains are much higher. Furthermore, Fig. 5 illustrates that irrespective of the V shape parameter each UE's individual throughput is achieved, demonstrating the effectiveness of preserving the slice isolation using the Lyapunov optimization.

Finally, to portray the trade-off between optimality and queue stability, we show the convergence time of the algorithm with respect to the V shape parameter in Fig. 6. The x-axis represents the V shape parameter, whereas the y-axis the convergence time in ms. As it can be shown in the figure, the convergence time increases with increasing V for all the scenarios. However, it decreases with increasing the number of BSs. For 2 BSs the convergence time of a single run consisting of 5000 slots, on average varies from 200 up to 1050 ms ranging the V parameter from 1 to 100. In other words, it takes approximately between 200 and 1050 slots, ranging the V parameter from 1 to 100 for the algorithm to obtain queue stability and to satisfy QoS requirements. Alternatively, for 5 BSs the average varies between 7 and 120 ms ranging V from 1 to 100. The intuition behind this result is that increasing the number of BSs increases the potential throughput when a careful RAN slicing algorithm is applied and as such the individual UEs requirements are achieved faster.

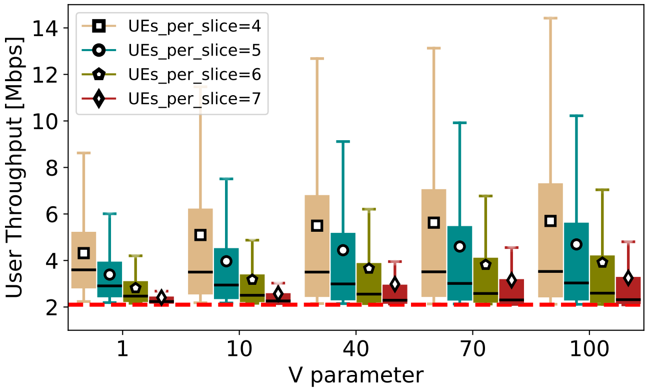


Figure 8: Individual UE throughput for Lyapunov optimization with increasing number of UEs for 2 BSs and 3 slices for various V shape parameters with a minimum requirement of 2.2 Mbps per UE. Lyapunov optimization can effectively preserve a minimum throughput for all UEs.

C. *Impact of number of UEs*

Similarly to the effect of the number of BSs to our algorithm, in this subsection we demonstrate the effect of UEs in the overall performance. By keeping the number of slices

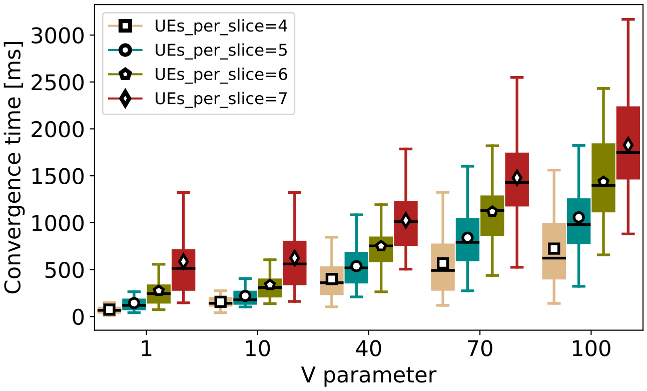


Figure 9: Converge time for Lyapunov optimization with increasing number of UEs for 2 BSs and 3 slices for various V shape parameters. The convergence time of the algorithm increases with increasing V and with number of UEs due to strict isolation constraints.

fixed to 3 and the number of BSs to 2, we show results with increasing number of UEs within a slice. Initially, we demonstrate the overall slice throughput, the slice isolation and finally the convergence time.

Fig. 7, demonstrates the overall slice throughput with respect to the number of UEs within a slice, whereas Fig. 8 the individual throughput achieved by each UE within a slice. The x-axis represents the V shape parameter, while the y-axis the overall slice throughput. Different boxplots are drawn for various UE sizes and the distribution of 50 runs is illustrated. To be noted that in this scenario for increasing number of UEs for 2 BSs the minimum requirement has been set to 2.2 Mbps per UE, otherwise the system is unable to provide the throughput due to infeasible solution. Again, all UEs have been combined together and the results are demonstrated in boxplots. In Fig. 7 we can observe that increasing the V shape parameter increases the overall slice throughput for all considered scenarios. The throughput ranges from 55 to 70 Mbps on average for 4 UEs per slice and 50 to 68 Mbps for 7 UEs per slice considering a V shaper parameter from 1 to 100. It is observed that increasing the number of UEs does not bring a big benefit in the overall slice performance. The intuition behind this result is greatly related to the UE minimum requirement that has to be preserved for each slice, which becomes challenging when more UEs exist in the system, especially since UEs are randomly selected from the aircraft plan in [47] and experience distinct channel variations. Nonetheless, Fig. 8 shows that for each V shape parameter the minimum requirement is achieved demonstrating the slice isolation.

Finally, we illustrate the trade-off between optimality and convergence time with respect to UEs in Fig. 9. The x-axis represents the V shape parameter, whereas the y-axis the convergence time in ms. As it is portrayed in the figure, the convergence time increases with increasing V as expected. Moreover, we notice that while the number of UEs increases the convergence time also increases. Considering 4 UEs per slice, the convergence time increases from 60 ms for $V = 1$ to 700 ms for $V = 100$ on average. Alternatively, for 7 UEs per slice, a variation between 600 ms for $V = 1$ and 1800 ms for $V = 100$ is experienced on average for the convergence time. This behavior can be explained with the ability to preserve each UE's minimum requirement, which is harder when the number of UEs increases in the network. However, as it was demonstrated above, the Lyapunov optimization is still able to achieve it. Again to be stressed, that the algorithm is used to solve the problem for a total of 5000 slots, thus the convergence time still remains within reasonable values.

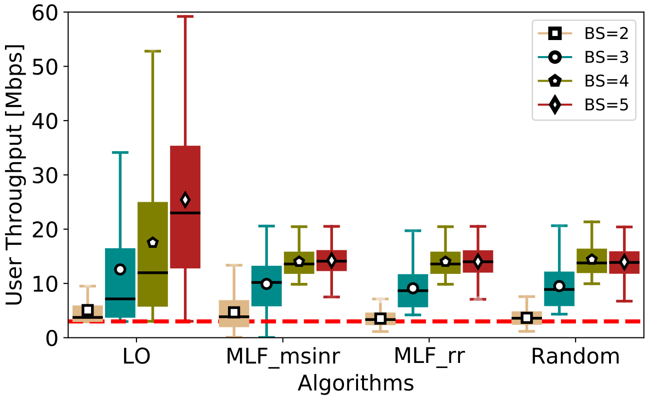
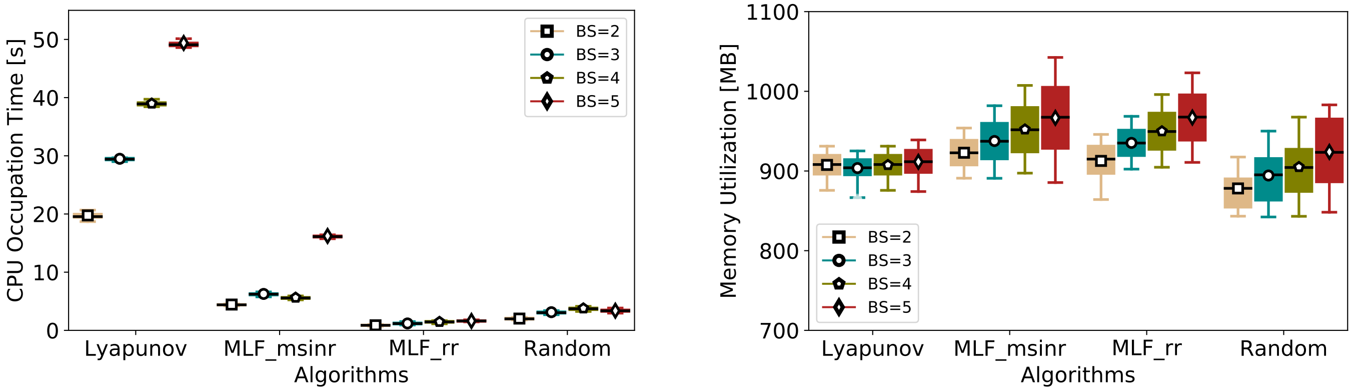


Figure 10: Comparison of Lyapunov optimization with state-of-the-art solutions for varying number of BSs for 3 slices and 12 UEs with minimum requirement 3 Mbps. The Lyapunov optimization can double the throughput compared to the best alternative solution MLF_msinr for 5 BSs. Moreover, it is the only technique that can guarantee all UEs' minimum requirement for all runs.

D. Algorithm Comparison Throughput

In this subsection, we demonstrate a comparison among two MLF [13] variants, our optimization approach as well as a random solution, where resources are randomly uniformly distributed over slices and users within a slice. Differently from Fig. 3, where we demonstrated the isolation comparison among Lyapunov and MLF variants only for a 2 BS scenario, here we investigate the difference also for higher number of BSs i.e., up to 5. We fix the number of UEs to 4 per each slice and we consider 3 slices. Setting up the minimum requirement for our approach to 3 Mbps per UE and V= 40, which demonstrates a good trade-off between optimality and convergence time, whereas for the alternative approach we consider a radio resource-based solution, where each slice requests equal resources. The comparisons are illustrated in Fig. 10, where the x-axis represents the different algorithms, whereas the y-axis the individual UE throughputs. The results are demonstrated for various BSs deployed in the network for 50 runs, where all UEs among slices are combined in single boxplots.

As we can observe by the figure, the Lyapunov optimization approach proposed in this work achieves slice isolation for all UEs compared to 50% of the best state-of-the-art solution i.e.,



(a) CPU utilization

(b) Memory utilization

Figure 11: Comparison of the Lyapunov optimization with state-of-the-art solutions for varying number of BSs for 3 slices and 12 UEs with minimum requirement 3 Mbps for V = 40 with respect to CPU and memory utilization. While the overhead for the memory utilization for the Lyapunov approach is minimal, the CPU consumption induces a higher overhead, but it is tolerable as compared to the state-of-the-art.

MLF maximum SINR. Moreover, it further achieves better overall performance with increasing BSs. Especially for 5 BSs the throughput is almost doubled on average, 26 Mbps compared to 14 Mbps of the best state-of-the-art solution. Again these results demonstrate the importance of considering a QoS based solution for frequency selective channels and indicate the alleviation in network performance i.e., throughput that can be accomplished with the Lyapunov approach.

E. Algorithm Comparison Overhead

In this subsection we elaborate more on the computational cost overhead of the proposed Lyapunov Optimization approach compared to the state-of-the-art. Similar to the previous comparison, there are in total 3 slices with 4 UEs each. The minimum requirement for our approach is set to 3 Mbps per UE with a V parameter of 40, which captures a good trade-off between optimality and convergence time. Results demonstrating the CPU and memory utilization of each algorithm for various BSs deployed in the network for 50 runs are illustrated in Fig. 11a and Fig. 11b, respectively. Fig. 11a shows the CPU occupation of each algorithm in seconds. According to the figure, the CPU utilization increases with increasing number of BSs. The Lyapunov optimization records the highest occupation time with up to 50 s for 5 BSs compared to the maximum SINR MLF solution with 18 s, demonstrating an induced overhead with respect to CPU consumption. Nonetheless, an improved code parallelization of the Lyapunov solution can further reduce the CPU overhead. Furthermore, we stress that the CPU consumption does not reflect the convergence time of the algorithm as shown in the previous results. Fig. 11b, illustrates a similar increasing trend of the memory utilization with increasing number of BSs for all the algorithms. However, the Lyapunov optimization records a smaller memory utilization on average apart from the random solution, concluding that the overhead of the Lyapunov optimization with respect to memory is minimal.

Finally, as mentioned in Section III, our approach is based on the SD-RAN paradigm, where the SD-RAN controller in combination with the SD-RAN agents i.e., BSs perform the resource allocation indicating a hierarchical scheduling approach. As an example framework for our solution the FlexRAN [26] SD-RAN platform can be utilized, as it offers the required architecture and protocol to achieve the communication among the SD-RAN controllers and the SD-RAN agents. Every coherence time τ i.e., 20 ms in our case, the BSs i.e., agents exchange UE information with the SD-RAN controller to track the latest changes of the wireless channel, which intuitively renders an overhead for such an approach. In order to demystify the aforementioned overhead, the work provided in [48] benchmarks the FlexRAN controller with respect to the communication overhead from the SD-RAN agents towards the SD-RAN controller. Considering the case of up to 5 BSs with 12 UEs the recorded overhead is less than 100 Kbit/s, which concludes that a per user basis hierarchical scheduling approach is not extremely demanding.

VI. Discussion

While we have demonstrated the effectiveness of the Lyapunov optimization in a multi-cell scenario compared to state-of-the-art approaches, in this subsection, we want to discuss the applicability of our algorithm in a realistic scenario. Initially, we detail how our algorithm can be implemented practically using open source platforms, whereas we explain whether the converge time of the algorithm is suitable for real time applications. Moreover, we discuss the generalization of our approach for larger scale scenarios as well as different use-cases and inputs for network slicing.

A. Practical implementation

In this work we focused on proposing a solution that can be adopted by standardization and that be practically realized. To that end, we follow the architecture proposed in [49] as well as the software-defined radio access network (SD-RAN) paradigm. The architecture presented in Fig. 2 is envisioned to be incorporated into existing SD-RAN open source platforms such as FlexRAN [26] combined with OpenAirInterface [50]. FlexRAN is designed such that it allows the communication with the application layer i.e., slice owners utilizing application programmable interfaces (APIs) similar to our approach. Moreover, FlexRAN provides a protocol that allows the communication with the underlying BSs and enables the possibility to alter and control resource allocation. Likewise, OpenAirInterface is a platform for 5G experimentation that has full compatibility with FlexRAN. The OpenAirInterface BS provides statistics about the UEs and the channel variations as well as the throughput achieved for each physical resource block (PRB). That entails, that this information can be utilized by our slice manager's statistics block in Fig. 2 to perform the optimization. Finally, the resource allocation is envisioned to operate within the FlexRAN controller in order to allow flexibility in the decision making with respect to the slice management. Although the integration of the proposed algorithm is not yet implemented in FlexRAN and OpenAirInterface, it remains an interesting research area in order to correlate the simulation results with a real implementation.

B. Convergence time

On the one hand, since resource allocation is a process that is performed every 1 ms, a proposed scheme shall be compliant with these requirements in order to be feasible in realistic scenarios. On the other hand, a resource allocation is dynamic for as long as the channel is dynamic i.e., wireless channel variations, user demand variations. In that regard a sophisticated resource allocation scheme shall always consider the coherence time of the channel i.e., the time the channel remains static and change the allocation when required. In our simulations, we have shown the convergence time of the algorithm considering a time horizon of 5000 slots with a coherence time of 20 ms. Our results demonstrate that the convergence time can reach on average up to 1050 ms for a time horizon of 5000 slots considering 2 BSs, but can decrease to 120 ms if 5 BSs are deployed. Nonetheless, in both cases the convergence time of the algorithm remains below the values of the time horizon it solves the problem for i.e., 5000 ms. That entails that our proposed scheme is feasible for realistic deployments.

C. Applicability in different scenarios

While in this paper we have considered up to 5 BSs and 21 UEs distributed evenly over 3 slices, the proposed scheme can be adjusted and solved for various number of BSs and UEs. It has been demonstrated that the Lyapunov optimization is benefiting from increasing the number of BSs not only in terms of throughput optimization, but also in terms of convergence time. Alternatively, increasing the number of UEs has an increasing effect on the convergence time. Nonetheless, the convergence time is always lower than the total time horizon considered i.e., 5000 slots as before mentioned.

Moreover, in our evaluation we have considered that all the network slices and the users of the slices have the same QoS requirements. However, our solution offers the possibility to consider every slice differently and therefore introduce a distinct QoS requirement per slice. This can be easily achieved by adjusting the values of constraint (9) depending on the slice's requirements in problem $\mathcal{P}_0$.

Finally, we would like to discuss the feasibility of the algorithm in terms of the minimum requirement per user. As aforementioned in Section V, for the Lyapunov optimization the minimum requirement per user is needed. This is typically a value requested by the slice owners. Nonetheless, it may occur that the environment is not able to satisfy all the requirements of the slices. In that case, an admission control mechanism becomes necessary. For our simulations, in order to demonstrate the effectiveness of our algorithm we have stressed the requirements to the channel limits in terms of throughput, yet keeping the solution feasible. In reality, this information can be acquired by utilizing estimation techniques for the channel throughput or by keeping a history of the records of the previously served users. Although this is an interesting research topic, it is not the scope of this paper, whereas we obtain this information by performing evaluations on the simulation platform.

VII. Conclusion

In this work, we consider the problem of RAN slicing in a multi-cell, frequency selective wireless environment. Considering the concept of softwarized RANs, we define a system framework to integrate our solution to existing SD-RAN platforms. Given many slice definitions in the state-of-the-art, we detail and identify that the correct definition of slice isolation should consider performance guarantees and not simply resource guarantees. Indeed, while in flat-fading channels the latter maybe sufficient, our results demonstrate that it will not be adequate for frequency selective channels. Moreover, we stress that a correct RAN slicing algorithm should cater for individual user requirements and not simply slice aggregate requirements. In that regard, we formulate the RAN slicing problem as a throughput maximization problem, while satisfying individual user performance guarantees. We have tackled the problem by using Lyapunov optimization that proves to be efficient in terms of maximizing slices' throughput and achieving slice isolation. In our results, we have demonstrated that in case that such a constraint is not introduced in the problem definition, individual user throughput degrades. We have performed extensive simulations to (i) demonstrate the importance of achieving a minimum required QoS performance for all users and (ii) efficiency with higher number of BSs due to channel gain considerations. Finally, we compare our approach with existing state-of-the-art solutions and show that it achieves better overall performance, while always guaranteeing users' requirements.

References

[1] E. J. Oughton, Z. Frias, S. van der Gaast, and R. van der Berg, "Assessing the capacity, coverage and cost of 5G infrastructure strategies: Analysis of the Netherlands," *Telematics and Informatics*, vol. 37, pp. 50–69, 2019.

[2] 3GPP, "TR 21.915 V1.1.0 (2019-03); Technical Report; Summary of Rel-15 Work Items (Release 15)," 2019.

[3] N. Alliance, “Description of network slicing concept,” *NGMN 5G P*, vol. 1, 2016.

[4] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega *et al.*, “Network slicing to enable scalability and flexibility in 5g mobile networks,” *IEEE Communications magazine*, vol. 55, no. 5, pp. 72–79, 2017.

[5] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, “Network slicing and softwarization: A survey on principles, enabling technologies, and solutions,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[6] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, “From network sharing to multi-tenancy: The 5G network slice broker,” *IEEE Communications Magazine*, vol. 54, no. 7, pp. 32–39, 2016.

[7] S. Mandelli, M. Andrews, S. Borst, and S. Klein, “Satisfying network slicing constraints via 5G MAC scheduling,” in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, 2019, pp. 2332–2340.

[8] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, “NVS: A substrate for virtualizing wireless resources in cellular networks,” *IEEE/ACM transactions on networking*, vol. 20, no. 5, pp. 1333–1346, 2011.

[9] Federal Communications Commission (FCC). Spectrum Crunch. [Online]. Available: https://www.fcc.gov/engineering-technology/policy-and-rules-division/general/radio-spectrum-allocation

[10] W. Nam, D. Bai, J. Lee, and I. Kang, “Advanced interference management for 5G cellular networks,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 52–60, 2014.

[11] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, “Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective,” *IEEE Wireless Communications*, vol. 21, no. 3, pp. 118–127, 2014.

[12] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson, “The role of small cells, coordinated multipoint, and massive MIMO in 5G,” *IEEE communications magazine*, vol. 52, no. 5, pp. 44–51, 2014.

[13] S. D’Oro, F. Restuccia, A. Talamonti, and T. Melodia, “The slice is served: Enforcing radio access network slicing in virtualized 5G systems,” in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, 2019, pp. 442–450.

[14] M. I. Kamel, L. B. Le, and A. Girard, “LTE wireless network virtualization: Dynamic slicing via flexible scheduling,” in *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*, pp. 1–5.

[15] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, “Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads,” *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 3044–3058, 2017.

[16] V. Sciancalepore, M. Di Renzo, and X. Costa-Perez, “STORNS: Stochastic radio access network slicing,” in *IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[17] C.-Y. Chang, N. Nikaein, and T. Spyropoulos, “Radio access network resource slicing for flexible service execution,” in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 668–673.

[18] A. T. Z. Kasgari and W. Saad, “Stochastic optimization and control framework for 5G network slicing with effective isolation,” in *52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018, pp. 1–6.

[19] A. Papa, M. Klugel, L. Goratti, T. Rasheed, and W. Kellerer, “Optimizing dynamic RAN slicing in programmable 5G networks,” in *IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[20] 3GPP, “Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.803, 2008. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2482

[21] H. G. Myung, J. Lim, and D. J. Goodman, “Single carrier fdma for uplink wireless transmission,” *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30–38, 2006.

[22] R. Ferrus, O. Sallent, J. Pérez-Romero, and R. Agusti, “On 5G radio access network slicing: Radio interface protocol features and configuration,” *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, 2018.

[23] A. Ksentini and N. Nikaein, “Toward enforcing network slicing on RAN: Flexibility and resources abstraction,” *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, 2017.

[24] J. Li, W. Shi, P. Yang, Q. Ye, X. S. Shen, X. Li, and J. Rao, “A hierarchical soft RAN slicing framework for differentiated service provisioning,” *IEEE Wireless Communications*, vol. 27, no. 6, pp. 90–97, 2020.

[25] X. Foukas, M. K. Marina, and K. Kontovasilis, “Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture,” in *Proceedings of the 23rd annual international conference on mobile computing and networking*, 2017, pp. 127–140.

[26] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, “FlexRAN: A flexible and programmable platform for software-defined radio access networks,” in *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*. ACM, 2016, pp. 427–441.

[27] E. Coronado, S. N. Khan, and R. Riggio, “5G-EmPOWER: A software-defined networking platform for 5G radio access networks,” *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 715–728, 2019.

[28] I. Vilà, J. Pérez-Romero, O. Sallent, and A. Umbert, “Characterization of radio access network slicing scenarios with 5G QoS provisioning,” *IEEE access*, vol. 8, pp. 51 414–51 430, 2020.

[29] R. Schmidt, C.-Y. Chang, and N. Nikaein, “Slice Scheduling with QoS-Guarantee Towards 5G,” in *IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–7.

[30] B. Han, J. Lianghai, and H. D. Schotten, “Slice as an evolutionary service: Genetic optimization for inter-slice resource management in 5G networks,” *IEEE Access*, vol. 6, pp. 33 137–33 147, 2018.

[31] 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. 55 765–55 779, 2018.

[32] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, “Optimising 5G infrastructure markets: The business of network slicing,” in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, 2017, pp. 1–9.

[33] J. X. Salvat, L. Zanzi, A. Garcia-Saavedra, V. Sciancalepore, and X. Costa-Perez, “Overbooking network slices through yield-driven end-to-end orchestration,” in *Proceedings of the 14th International Conference on emerging Networking EXperiments and Technologies*, 2018, pp. 353–365.

[34] J. Mei, X. Wang, K. Zheng, G. Boudreau, A. B. Sediq, and H. Abouzeid, “Intelligent Radio Access Network Slicing for Service Provisioning in 6G: A Hierarchical Deep Reinforcement Learning Approach,” *IEEE Transactions on Communications*, 2021.

[35] D. Ginthör, R. Guillaume, M. Schüngel, and H. D. Schotten, “5G RAN Slicing for Deterministic Traffic,” in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1–6.

[36] W. Shi, J. Li, P. Yang, Q. Ye, W. Zhuang, S. Shen, and X. Li, “Two-level Soft RAN Slicing for Customized Services in 5G-and-beyond Wireless Communications,” *IEEE Transactions on Industrial Informatics*, 2021.

[37] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, “On radio access network slicing from a radio resource management perspective,” *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, 2017.

[38] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, “Network slicing games: Enabling customization in multi-tenant mobile networks,” *IEEE/ACM Transactions on Networking*, vol. 27, no. 2, pp. 662–675, 2019.

[39] Y. Sun, S. Qin, G. Feng, L. Zhang, and M. Imran, “Service provisioning framework for RAN slicing: user admissibility, slice association and bandwidth allocation,” *IEEE Transactions on Mobile Computing*, 2020.

[40] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.

[41] N. Moraitis and P. Constantinou, “Radio channel measurements and characterization inside aircrafts for in-cabin wireless networks,” in *IEEE 68th Vehicular Technology Conference*, 2008, pp. 1–5.

[42] M. J. Neely, “Stochastic network optimization with application to communication and queueing systems,” *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[43] C. Wallace. (2019) Bringing 5G networks indoors. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/bringing-5g-networks-indoors

[44] Huawei. (2016) Small cell network. [Online]. Available: https://www.huawei.com/minisite/hwmbbf16/insights/small_cell_solution.pdf

[45] Nokia. (2017) Small cells and femtocells. [Online]. Available: https://www.nokia.com/networks/portfolio/small-cells/#small-cells

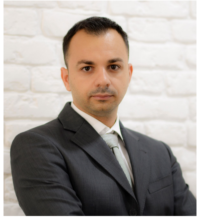
[46] N. Moraitis, P. Constantinou, F. Perez Fontan, and P. Valtr, “Propagation measurements and comparison with EM techniques for in-cabin wireless networks,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1–13.

[47] The boeing, 2007. [Online]. Available: http://www.boeing.com

[48] A. Papa, R. Durner, E. Goshi, L. Gorattiy, T. Rasheedy, A. Blenk, and W. Kellerer, "MARC: On Modeling and Analysis of Software-Defined Radio Access Network Controllers," *IEEE Transactions on Network and Service Management*, 2021.

[49] A. Papa, R. Durner, L. Goratti, T. Rasheed, and W. Kellerer, "Controlling Next-Generation Software-Defined RANs," *IEEE Communications Magazine*, vol. 58, no. 7, pp. 58–64, 2020.

[50] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.



Arled Papa completed his Bachelor of Science in Electronics Engineering at the Polytechnic University of Tirana, Albania in 2015. He received his Master of Science in Communications Engineering at the Technical University of Munich (TUM) in November 2017 with high distinction. In February 2018 he joined the Chair of Communication Networks at TUM as a research and teaching associate. His research focuses on the design and analysis of QoS, Network Slicing and Programmability of Sofware-Defined Radio Access Networks.

Alba Jano received her Bachelor of Science degree in Telecommunications Engineering at Polytechnic University of Tirana, Albania, in July 2017. She completed her Master of Science in Communications Engineering at the Technical University of Munich (TUM) in January 2020 with distinction. In February 2020, she joined the Chair of Communication Networks at TUM as a research and teaching associate. Her research focuses on the design and evaluation of methods for wireless resource management in 5G and beyond networks and for digital twins' communication.

Serkut Ayvaşık received his Bachelor of Science degree in Electrical and Electronics Engineering at Middle East Technical University (METU), Turkey in 2016. He obtained the M.Sc. Degree in Communications Engineering in February 2019 with high distinction from Technical University of Munich (TUM) as a scholarship holder of German Academic Exchange Service (DAAD) and Turkish Education Foundation (TEV). In March 2019, he joined the Chair of Communications Networks at TUM as a research and teaching associate. His research interests are design and evaluation of radio resource management for URLLC applications and machine learning in 5G/6G wireless networks.

Onur Ayan received his B. Sc. and M. Sc. degrees in electrical and computer engineering at Technical University of Munich. In January 2017 he joined the Chair of Communication Networks where he is currently a doctoral candidate and working as a teaching associate. His current research interests include design and evaluation of wireless networked control systems, network support for cyber-physical systems and age of information.

H. Murat Gürsu (S16) was born in Istanbul, Turkey, in 1989. He received the B.Sc. degree in electrical and electronics engineering from Bogazici University in 2012, the M.Sc. degree in communication engineering from the Technical University of Munich (TUM) in 2014 and his Ph.D. degree where his thesis has the title Delay-Constrained Reliable Cellular Uplink Radio Resource Management for Industrial Internet of Things in 2020 from the Chair of Communication Networks, Department of Electrical and Computer Engineering in TU Munich. Currently, he is working as a Radio Access Research Specialist in Nokia Bell Labs on 5G standardization.

Wolfgang Kellerer (M'96, SM'11) received the Dr.-Ing. (Ph.D.) and Dipl.-Ing. (Master) degrees from the Technical University of Munich, Germany, in 1995 and 2002, respectively, where he is a Full Professor, heading the Chair of Communication Networks with the Department of Electrical and Computer Engineering. He was with NTT DOCOMOs European Research Laboratories for ten years in leading positions, contributing to research and standardization of LTE-A and 5G technologies. In 2001, he was a Visiting Researcher with the Information Systems Laboratory, Stanford University, CA, USA. His research has resulted in over 200 publications and 35 granted patents. He was awarded with an ERC Consolidator Grant from the European Commission for his research project FlexNets Quantifying Flexibility in Communication Networks in 2015. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and as an area editor for network virtualization for IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He is a member of ACM and VDE ITG.