IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 10, 2020, accepted April 21, 2020, date of publication April 27, 2020, date of current version May 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990802

Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond

PABLO MUÑOZ [1], ÓSCAR ADAMUZ-HINOJOSA [1], JORGE NAVARRO-ORTIZ [1], ORIOL SALLENT [2], AND JORDI PÉREZ-ROMERO [2], (Member, IEEE)
[1] Department of Signal Theory, Telematics and Communications, Universidad de Granada, 18071 Granada, Spain
[2] Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain

Corresponding author: Pablo Muñoz (pabloml@ugr.es)

This work was supported in part by the H2020 Research and Innovation Project 5G-CLARITY under Grant 871428, in part by the Andalusian Knowledge Agency under Project A-TIC-241-UGR18, in part by the Spanish Ministry of Education, Culture and Sport (FPU) under Grant 17/01844, and in part by the Spanish Research Council and FEDER funds through SONAR 5G under Grant TEC2017-82651-R.

ABSTRACT The new fifth generation (5G) era has transformed previous mobile generations into fast, smart networks that will be more responsive and customizable. With network slicing, 5G networks can be dynamically adapted to the different needs of specific vertical industries. This capability has opened the opportunity to new business models whereby infrastructure owners can monetize their investment by leasing resources to third parties (i.e. tenants). In this respect, a challenging task for the owner of the radio access network infrastructure (i.e. the network provider) is the spectrum planning of multi-tenant scenarios. This paper proposes different alternatives of hiring capacity to the provider as well as a set of spectrum planning strategies, each giving a certain degree of flexibility to allocate resources per tenant. These strategies are evaluated in a 5G small cell multi-tenant network through snapshot-based simulations. The performance of the strategies is assessed in terms of scalability, spectrum isolation, utilization and efficiency.

INDEX TERMS 5G networks, multi-tenancy, new radio, RAN slicing, small cells, spectrum planning.

I. INTRODUCTION

The enormous growth of subscribers' data traffic in the last years has stressed the need of a substantial change on current mobile networks. Likewise, the growing industrial digitalization has boosted a wide range of novel applications with stringent and business critical requirements. To meet these rising and diverse demands, the 5th generation (5G) mobile network has introduced innovative architectural and technological features [1], such as network slicing, network softwarization, massive multiple-input multiple-output (MIMO) and device-to-device communications.

Currently, the cost of upgrading the infrastructure is extraordinary for most mobile network operators (MNOs) since they rely on relatively low-cost flat rates. Deploying new infrastructure entails large delays due to site acquisition and installation, spectrum leasing, etc. Furthermore, in an operational network, there are underutilized resources due to traffic demand variations. Under these premises, the traditional scenario with independently deployed networks is unfeasible to embrace the 5G network evolution. Instead, to provide cost-efficient solutions with a shorter time-to-market, new business models based on cooperation and infrastructure sharing are needed [2]. In this way, services can be deployed faster while reducing capital (CAPEX) and operational expenditures (OPEX).

Network sharing is a paradigm that enables MNOs to act as infrastructure providers, leasing the infrastructure to other MNOs or mobile virtual network operators (MVNOs) for entering the market or extending coverage/capacity. Multi-tenancy is an extension of this concept where a third-party making use of the infrastructure as a tenant becomes a service provider, such as those offering over-the-top (OTT) applications (e.g. streaming) or vertical industries (e.g. manufacturing, entertainment, public safety) [3].

Service providers or tenants impose diverse technical and business requirements to the network. To provide efficient deployment of these services, the network should be flexible and scalable. Network slicing has been proposed as an efficient solution to provide flexibility and scalability in the 5G mobile networks [4], [5]. This feature consists in creating multiple logical, self-contained networks on top of a common

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

shared physical infrastructure and, therefore, it can be used to support multi-tenancy on the 5G network. In this case, each network slice is specifically built to meet the service requirements of a certain tenant (e.g. in terms of speed, capacity, connectivity and coverage). Technologies such as Software-Defined Networks (SDN) and Network Function Virtualization (NFV) are key enablers to the implementation of network slicing [6]. They enable the use of common resources such as storage and processors to run logical (software-based) elements that can be controlled programmatically.

Network slicing also provides adequate resource isolation, independent scaling and increased statistical multiplexing. Creating an independent virtualized end-to-end (E2E) network involves the configuration of the radio access network (RAN), transport, and core network [7]. However, the complexity of the configuration in the RAN is greater due to difficulties in partitioning radio resources and virtualizing functionalities with tight latency requirements [8]–[9]. On the one hand, slices cannot interfere with each other to ensure isolation. A strict resource isolation implies orthogonal spectrum allocation between slices. This may result in inefficient resource utilization, especially in large service areas with varying traffic demands. On the other hand, the lower layers of the radio protocol stack have a large number of interfaces and varying capabilities that operate on a very fast timescale. This certainly complicates the virtualization and limits the functional split options between a Centralized Unit (CU) and distributed unit (DU) in a 5G RAN node.

The benefits of network slicing in the 5G RAN, or Next Generation RAN (NG-RAN), will rely on the flexibility and scalability offered by the lower layers of the radio protocol stack. In this way, the Third Generation Partnership Project (3GPP) has defined service and operational requirements for 5G network slicing [10] and technical specifications for the 5G air interface, known as New Radio (NR) [11]. The latter includes key technology features in the physical layer such as scalable numerology to support multiple bandwidths and spectrum and flexible frame structure to provide low latency and high efficiency. Thus, the high degree of configurability offered by the NR enables better resource sharing between tenants and better customization of slices according to service requirements.

In the 5G-RAN, the spectrum planning is in charge of allocating spectrum resources to each slice before its operation based on capacity and isolation requirements. There can be different ways to perform spectrum planning depending on the service-level agreement (SLA) between the network provider and tenant. This SLA determines how the provider can allocate spectrum resources over the network for each slice. Depending on the required level of isolation, for example, a slice can require exclusive (i.e. non-shared) use of a resource in the entire network or, alternatively, the resource can be shared in different cells with other slices, so that exclusive use is limited to the cell area. Each level of radio isolation determines the multiplexing gains and gives the provider some degree of flexibility for allocating spectrum resources. In this way, a slice with exclusive use of radio resources will prevent other slices from using them. The impact of the radio isolation on the spectrum planning have not been analyzed yet with the required depth.

In this work, we propose different ways of hiring capacity to the network provider and, based on them, we analyze a set of spectrum planning strategies with different degrees of flexibility for allocating spectrum resources. The proposed strategies are evaluated in terms of scalability, spectrum isolation, utilization and efficiency in a 5G small cell (SC) network. SCs can help satisfy the increasing traffic demand while they facilitate the adoption of network slicing [12]. However, these low-power devices entail more complex spectrum planning than macro-cells because they facilitate extensive spatial reuse.

The remainder of this paper is organized as follows. In Section II, the literature related to RAN slicing is discussed. Section III describes the system model. In Section IV, the proposed strategies for spectrum planning are presented. Section V provides the performance results for the different strategies. Finally, Section VI summarizes the conclusions.

II. RELATED WORK

The importance of network slicing has been widely recognized, becoming a fundamental topic in many research initiatives. Diverse standard organizations such as 3GPP, ETSI, ITU and IETF are spending much effort to network slicing, offering different views of it. There is a broad consensus that the SDN/NFV paradigm is a key enabler to provide functional customization over the same infrastructure. Comprehensive reviews on SDN/NFV-based solutions related to network slicing are discussed in [13], [14]. The work in [15] analyzes a proposal from ETSI that incorporates the capabilities of SDN into the NFV architecture to enable the realization of network slices. In [16], a slicing-enabled SDN core network architecture is proposed for the automotive vertical use case. From the management viewpoint, the work in [17] proposes a SDN/NFV-based framework to manage E2E network slices, including their lifecycle and context management, monitoring and configuration. The creation process of network slices is addressed in [18], where network slice descriptors are used to make this process more agile and automatic. The idea of a network and application store is introduced in [19] to simplify the procedure of defining the network slice. It provides a marketplace for delivering customized network functions and service templates tailored to specific use cases.

RAN slicing [20] poses many interesting challenges related to the management of the slice's lifecycle, as well as the abstraction and sharing of radio resources. In [21], the issue of RAN resource allocation is addressed considering resources of a base station such as radio bandwidth, caching, and backhaul components. The support for latency-sensitive and time-critical applications through RAN slicing is investigated in [22], where the number of radio resources and their relative position in the time domain are considered to satisfy the latency requirements. Similarly, the work in [23] presents a novel slice resource allocation approach that introduces the concept of mini-slots to support low latency communications. The issue of spectrum allocation to minimize inter-slice interference is analyzed in [24], where various algorithms are developed to guarantee orthogonality among RAN slices. However, such orthogonality may lead to inefficient resource usage. With a special focus on the E2E isolation, a systematic overview of existing isolation techniques is provided in [25]. Nevertheless, an exhaustive analysis of the radio isolation is still missing.

The concept of RAN slicing at different levels is introduced in [26], where each defined level provides a specific degree of granularity in the assignment of radio resources, isolation and customization. The spectrum allocation at the scheduler level is investigated in several works [27], [28]. The scope of these works lies on the dynamic resource allocation (operating at a faster time scale than the spectrum planning level) to cope with the traffic dynamics. The dynamic slice scheduling using a centralized approach (i.e. a SDN-enabled controller) for heterogeneous networks has been addressed in [29]. In [9], the link-layer scheduler is partitioned into two levels to perform inter- and intra-slice scheduling. In [30], the two-level hierarchy is implemented by giving priority to the different slices (e.g. prioritizing enhanced mobile broadband) and the users within the slices. In [31], this hierarchy is internal to the base station and supported by a centralized entity that controls spectrum sharing between tenants. At the spectrum planning level, the work in [32] proposes a spectrum planning scheme that maximizes the spectrum utilization. However, the isolation issue is not considered as part of the optimization problem.

Artificial intelligence (AI) has also been successfully applied to network slicing. In [33], an AI-enabled 5G network architecture is proposed to adjust service configuration and control based on changes in user needs, environmental conditions and business goals. Some interesting AI techniques that have recently been applied to resource allocation of slices are reinforcement learning (RL) [34], deep RL [35]–[38], deep learning neural networks [39], [40] and evolutionary algorithms [41]. These techniques are particularly effective in handling complicated control problems.

There are still few works analyzing the impact of the realization of RAN slicing from a management perspective. In [42], a framework is proposed for the specification of RAN slices based on a set of configuration descriptors that characterize features, policies and resources. Such a framework has been extended in [43] with the specification of certain radio resource management functionalities (e.g. admission control and packet scheduler) as part of the RAN slice configuration.

Although there have been substantial research on RAN slicing, there is still little work on analyzing in detail different business-driven models for multi-slice/tenant spectrum sharing. The work in [44] proposes two spectrum sharing models and algorithms with different level of flexibility, depending on whether a set of dedicated and shared resources per tenant are predefined or not. However, the algorithms assign spectrum resources per user, acting as a link-layer scheduler. Thus, this algorithm may not scale well in large networks, as traffic demand variations among cells are not considered. In addition, it is hard to measure whether the offered capacity conforms or not to the contract because it would require extensive metric monitoring in the network. The present work tackles the problem of spectrum sharing at the planning level, which provides a wider view of the network, enabling optimal resource allocation per cell. In addition, this level facilitates the mapping of high-level capacity specifications to lower-level constraints, ensuring fairer resource allocation among tenants.

This work further develops the functional framework proposed in [45], [46] to include isolation as a key feature for slice specification and propose appropriate business-driven models for spectrum sharing. Such a framework enables self-planning of the radio access capacity in a NG-RAN, including automatic cell re-configuration mechanisms in order to facilitate the realization of slices. Under this framework, the contributions and novelties of this paper are the following:

1) Proposing an effective business-driven model for capacity specification, simplifying the later capacity compliance analysis.
2) Introducing isolation as a part of the slice specification for RAN slices to ensure that the traffic load of one slice does not negatively affect other slices.
3) Proposing different spectrum sharing strategies at the planning level for RAN slicing, each giving a certain degree of flexibility to allocate resources per slice.
4) Providing an exhaustive analysis of the proposed strategies in terms of scalability, spectrum isolation, utilization and efficiency in a NG-RAN. In essence, there is a trade-off between maximizing resource usage and avoiding/reducing co-channel interference between slices.

III. SYSTEM MODEL

A. NETWORK MODEL

Consider a NG-RAN consisting of a set B of 5G NR SCs that are owned by a certain infrastructure provider. The SCs are conceived to satisfy high traffic demands in localized areas. Multiple tenants (e.g. OTT providers or industry vertical market players) can request and lease resources from the infrastructure provider to deploy a set S of network slices. The slice s provides a service over a certain area specified through a subset $B_s \subseteq B$ of SCs. The aggregated service demand D of the slices is non-uniformly distributed over the considered area. Accordingly, the network topology is assumed irregular (i.e. with cell service areas of different size) to absorb the service demand with maximum resource usage efficiency. Under this demand distribution, a set U of user equipments (UEs) exist in the scenario, being U_s, $s \in S$ the subset of UEs belonging to slice s.

The UEs of a slice should be provided with enough resources to satisfy a guaranteed bit rate or service demand.

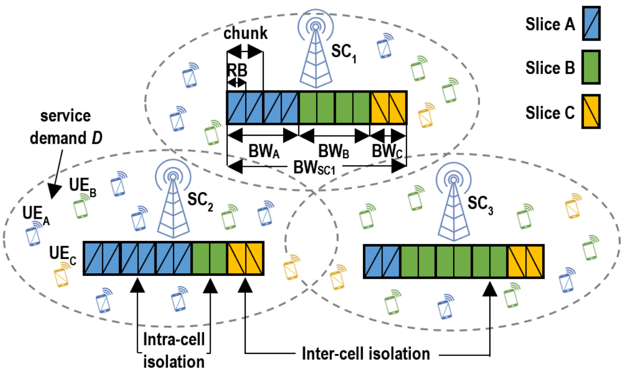


FIGURE 1. Resource model of the NG-RAN and an example with different levels of isolation among slices.

In particular, the accessing scheme is orthogonal frequency division multiple access (OFDMA), since 3GPP agreed to adopt it for 5G NR [11]. Specifically, the system supports scalable numerologies with subcarrier spacing of $2^{\mu}\cdot 15$ kHz ($\mu = 0,1,\ldots,4$). As shown in Fig. 1, the system bandwidth (BW) is divided into a set of resource blocks (RBs), each consisting of 12 consecutive subcarriers in the frequency domain. In Release 16 [47], the number of RBs ranges from 11 to 273 units. Depending on the numerology, the number of RBs is mapped to a specific bandwidth. For example, a single NR carrier with 133 RBs would require 25 MHz bandwidth for $\mu = 0$ or 50 MHz for $\mu = 1$. The maximum allowable bandwidth depends on the spectrum band where NR operates. In particular, this limit is 100 MHz for the sub-6 GHz band and 400 MHz for the millimeter wave band. At frequencies below 6 GHz, the cell size is larger and subcarrier spacing of 15 and 30 kHz are appropriate, while at higher frequencies, subcarrier spacing of 60 and 120 kHz are more suitable.

The number of RBs is typically high for most system bandwidths. Then, from the perspective of allocating the spectrum resources to the different slices, it becomes advantageous to reduce the management complexity by grouping the RBs into spectrum chunks, which are allocated to the slices as an indivisible unit. This can be done through the concepts of bandwidth part and RB group defined in [48] and [49], respectively. The bandwidth part is a subset of contiguous common RBs for a given numerology. This new feature will enable the coexistence of multiple slices with different physical layer requirements. The RB group is a collection of RBs within a given bandwidth part that can be allocated to the scheduled UEs. The size of the spectrum chunk can be used to establish the minimum allocation unit size. This parameter may serve to reduce the signaling overhead at the expense of a loss of flexibility, which could be critical when the number of slices is large.

Bearing in mind these considerations, the RBs are grouped into a set R of spectrum chunks (see Fig. 1). Each chunk is composed of a number of RBs equal to the minimum allocation unit size. From the set of chunks, a subset R_b of the available chunks in the system bandwidth is allocated to the SC b, $b \in B$. In addition, among these chunks, $R_{b,s}$ are allocated to the slice s, $s \in S$. Depending on how these chunks are allocated to the slices, intra-cell or inter-cell isolation can be provided as will be explained in detail in the next section. In any case, to provide a slice with full coverage within its service area B_s, at least one chunk is allocated in every SC, i.e. $|R_{b,s}| > 0$, $b \in B_s$.

The subset of R_b allocated chunks provides the SC bandwidth BW_b, which in turn determines the required transmit power, P_b^{TX} of the SC b. In particular, the transmit power must ensure a targeted Signal-to-Interference-plus-Noise Ratio (SINR) at the cell coverage range, i.e.:

$$P_b^{TX} = \min(P_N \cdot G_{PL,b}(d_{edge}) \cdot BW_b \cdot SINR_{edge}, P_{max}^{TX}), \quad (1)$$

where P_N is the noise power measured in one chunk, $G_{PL,b}(d_{edge})$ is the path gain (loss) evaluated at the distance d_{edge} between the SC and the cell-edge, $SINR_{edge}$ is the target value at that distance and P_{max}^{TX} is the maximum transmit power. The cell-edge is determined by the distance to the closest adjacent SC.

The received power $P_b^{RX}(d)$ at a certain distance d when served by the SC b is given by:

$$P_b^{RX}(d) = P_b^{TX} \cdot G_b(d), \quad (2)$$

where $G_b(d)$ is the overall gain at the distance d including the antenna gain, the shadow fading (loss) and the path loss. The fast fading is not modelled as the channel gain is measured over a large time scale.

The $SINR(u,r)$ experienced by the UE u when transmitting on the chunk r is defined as:

$$SINR(u,r) = \frac{P_b^{RX}(d_{b,u})}{\left(\sum_{j\in B\setminus\{b\}} L_j \cdot \pi_j(r) \cdot P_j^{RX}(d_{j,u})\right) + P_N}, \quad (3)$$

where $d_{b,u}$ is the distance between the SC b and the UE u, L_j is the cell load factor of the SC j and $\pi_j(r)$ is a function that takes the value 1 when the chunk r is allocated to the SC j and the value 0 otherwise. The cell load factor is determined from the relation between the service demand and cell capacity, i.e.:

$$\hat{L}_j = \frac{\sum_{u|j=\Gamma(u)} D_u}{\sum_{u|j=\Gamma(u)} BW_u \cdot SE_u} \quad (4)$$

and

$$L_j = \min\left(\hat{L}_j, 1\right), \quad (5)$$

where D_u and SE_u are the service demand and spectral efficiency of the UE u, respectively, BW_u is the fraction of the cell bandwidth allocated to this UE according to the slice's constraints and the resource scheduling policy and $\Gamma(u)$ is a function that returns the serving SC based on the strongest SINR. Then, the cell overload factor for the SC j is defined as:

$$OL_j = \hat{L}_j - L_j. \quad (6)$$

This variable serves as an indicator of the congestion level in the network. Thus, under congested situations, it will be greater than zero.

The spectral efficiency $SE(u,r)$ of the UE u in the chunk r is derived from the $SINR(u,r)$ according to the following SINR mapping [50]:

$$SE = \begin{cases} 0, & SINR < SINR_{min} \\ \alpha \cdot \log_2(1 + SINR), & SINR_{min} \leq SINR < SINR_{max} \\ SE_{max}, & SINR \geq SINR_{max}, \end{cases} \tag{7}$$

where SE_{max} is the maximum achievable spectral efficiency with link adaptation, $SINR_{min}$ and $SINR_{max}$ are the minimum and maximum SINR values, respectively, and α stands for the attenuation factor, which represents implementation losses. Lastly, the UE throughput $T(u)$ is given by:

$$T(u) = \min\left(\frac{BW_u}{|R_{b,s}|} \cdot \sum_{r \in R_{b,s}} SE(u,r), D_u\right), \tag{8}$$

where $b = \Gamma(u)$ is the serving SC. The UE throughput depends on the resource scheduling scheme through the variable BW_u. For example, assuming a Round-Robin scheme, this variable is given by:

$$BW_u = \frac{BW_{ch} \cdot |R_{b,s}|}{|U_{b,s}|}, \tag{9}$$

where BW_{ch} is the bandwidth of one chunk and $U_{b,s}$ stands for the subset of UEs connecting to the same SC b and slice s that fairly share the spectrum, i.e. $U_{b,s} = \{v | v \in U_s \land b = \Gamma(v)\}$.

B. RAN SLICING FRAMEWORK

A RAN slice defines a particular behavior of the NG-RAN in terms of capabilities and parameters configuration to meet the service requirements specified by the tenant. A key aspect in the orchestration and configuration of the RAN is how the radio spectrum is allocated and shared between the slices. The infrastructure provider is responsible for deploying and operating a number of concurrent RAN slices, including procedures such as slice instantiation, scaling and termination. These procedures should be carried out in an automated and agile way, allowing rapid adaptation to the business needs.

In order to address these issues, the general framework for network slicing with focus on the RAN is depicted in Fig. 2. This framework is based on a layered architectural approach and it is well aligned with most proposals from the literature [13], [27], and [46], as well as standardization bodies (e.g. 3GPP, ETSI). Going into details, the service layer acts as the interface between the tenant and the infrastructure provider through a set of management functions to support several tasks such as SLA negotiation or performance monitoring. To describe the service requirements with a high abstraction level, the tenants has at its disposal the Generic Slice Template (GST) [51], which is a set of attributes that characterize a type of network slice (e.g. a mobile broadband slice). The tenant is asked to modify the GST to include its particular service requirements, giving place to the generation of the Individual Slice Template (IST). This template may contain requirements from the SLA, e.g. key performance indicators for throughput or latency, and other aspects such as demand patterns and additional services. The service layer delivers the IST containing the service level description to the management layer. This latter comprises the Network Slice Management Function (NSMF), which is responsible for the creation and operation of the E2E slice. The NSMF relies on several Network Slice Subnet Management Functions (NSSMFs), each of which covers a particular network domain (e.g. access or core). To reduce the operational complexity, each resource domain that constitutes the E2E slice may have its own IST. Consequently, the RAN IST allows customizing the functions, policies and resources within the radio protocol stack for the RAN slice configuration. The RAN NSSMF is in charge of translating the slice requirements included within the IST to the configuration parameters in the RAN. To achieve this, the RAN NSSMF may support a wide range of internal functions for provisioning and performance/fault management, such as spectrum planning, admission control, SLA conformance monitoring or traffic forecasting. Among them, the spectrum planning manages the long-term allocation of spectrum chunks for each slice given its capacity requirement and the desired level of resource isolation.

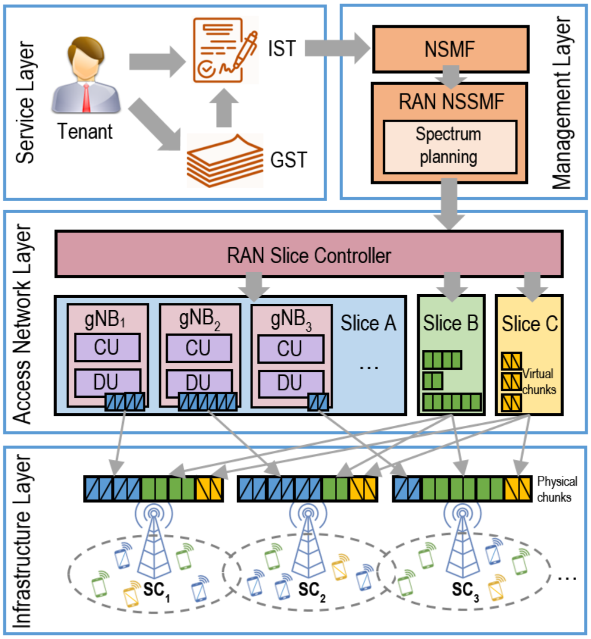


FIGURE 2. Architectural framework for network slicing.

The access network layer comprises multiple instances of different combinations of logical resources grouped as slices. The main logical unit in the RAN is the Next-Generation NodeB (gNB), which hosts the full radio protocol stack functionality and it is decomposed into the CU and DU. This logical division provides deployment flexibility to split and move NR functions between the CU and DU entities. There is a broad consensus on the suitability of the SDN/NFV framework to implement these network functions. The logical slices are managed by a programmable RAN slice controller that is responsible for short-term decisions (e.g. inter-slice scheduling) considering the traffic dynamics of the slices and the guidance, parameters and policies provided by the RAN NSSMF.

Lastly, the infrastructure layer comprises the set of physical network resources in the RAN including SCs, edge data centers and interconnections through fiber or wireless-based transport networks. The RAN resources are located in strategic Points of Presence, e.g. the cell sites for SCs and the central offices for edge data centers. The virtualized resources (e.g. a certain number of virtual chunks) need to be mapped to actual physical resources (e.g. a set of frequencies in the spectrum band). To facilitate this task, the 5G NR physical layer incorporates new features such as the bandwidth part [48].

IV. SPECTRUM PLANNING STRATEGIES FOR RAN SLICING SYSTEM MODEL

The first part of this section introduces a business-driven model to specify network capacity requirements. Then, it analyzes how these specifications are translated into different spectrum allocation strategies.

A. SLICE SPECIFICATION

The description of the slice requirements made by the tenant involves, besides quantitative definitions of the required capacity, a set of conditions under which the leased capacity is operated and managed (e.g. the isolation). This specification should be simple and expressed with a high abstraction level, avoiding details that might complicate other management tasks, such as the capacity conformance testing or the bandwidth throttling. With the aim of automating this procedure, and following the same principles as in Fig. 2, the tenant is invited to build an IST containing the necessary parameters and their configuration for the capacity provisioning.

Specifically, the proposed IST should include the following attributes or parameters that are related to the required capacity and isolation (i.e. application-specific attributes are out of the scope of this paper).

1) REQUESTED CAPACITY VALUE(S)

It specifies the capacity value(s) $v^{(s)}$ that satisfies the requirements of the slice s. This parameter can accept multiple definitions or values according to other parameters. For example, depending on the link direction, it can be downlink or uplink capacity; depending on the QoS class, it can be maximum or guaranteed capacity. It can also depend on the reservation type or the granularity level. Since these latter characteristics deserve special attention, they will discussed with more details below. In any case, the infrastructure provider must guarantee the requested capacity value(s) under the conditions given by the additional parameters.

2) CAPACITY RESERVATION TYPE

It determines whether the requested capacity is expressed in terms of throughput (e.g. in Mbps) or in terms of number of resources (e.g. in chunk units). While the former entails more accuracy in specifying the required capacity at the service layer, the latter can simplify the SLA compliance analysis and provide a fair resource usage between slices, especially under strong interference and congestion situations. In case of a throughput-based specification, the capacity conformance testing is carried out at the service layer, regardless of the underlying RB utilization. This may result in unfair resource sharing for low-traffic slices, where the UEs are receiving higher interference from heavily loaded slices. The choice on the reservation type may introduce additional parameters. For example, a throughput-based specification could also include a capacity margin to accept some excess traffic while there are available resources.

3) CAPACITY GRANULARITY LEVEL

Besides the various definitions of capacity already explained, the requested capacity value(s) can also be defined with a certain granularity level in the network. In particular, this parameter specifies whether the requested capacity value is defined on a per-UE (i.e. the service demand), per-cell or per-network basis. The per-UE capacity represents the lowest level of granularity. In case the requested capacity is only given per-UE, the maximum number of sessions or the maximum number of UEs (either per cell or per network) should be specified to quantify the aggregated capacity that is required in the network. The per-cell capacity implies guaranteeing an exact amount of capacity in every SC. Lastly, the per-network capacity enables a more flexible capacity allocation among SCs according to the spatial traffic distribution. In this case, 'network' refers to the cluster of SCs, B_s, which provide the service. Note that the per-network capacity requirement is equivalent to targeting an average cell capacity, calculated over the number of SCs in the cluster.

4) RESOURCE ISOLATION LEVEL

This parameter indicates the degree of resource isolation with other slices. Consider the chunk i to be allocated in the SC j for a given slice. The following isolation levels are distinguished: (i) no isolation: other slices can use the chunk i in any SC, including the SC j; (ii) intra-cell isolation: other slices can use the chunk i in a SC other than SC j; and (iii) inter-cell isolation: other slices cannot use this chunk in the entire network, ensuring radio-electrical isolation. Introducing isolation may simplify capacity management and supervision. For example, due to the individual usage of resources, capacity conformance testing would not be necessary. With no isolation, accounting for the resource consumption over the time is required for ensuring equal RB sharing between slices or, alternatively, the packet scheduler could adopt a fair-share scheduling policy.

5) SPATIO-TEMPORAL CONSTRAINTS

The spatial constraints limit the extent of the service area (e.g. a factory, a stadium, a mall, etc.) and, therefore, the cluster of SCs, B_s, serving the UEs. The time constraints define the time window(s) during which the service is offered to the UEs. The requested capacity value could be dependent on the spatial and time domains (e.g. to assign greater values in peak periods or high-traffic areas) at the expense of a more complex management.

From an economic viewpoint, the resource-based specification with intra- or inter-cell isolation enables a fairer sharing model than the throughput-based specification, since the tenants are charged based on their resource consumption while the inter-slice interference can be avoided or limited. Additionally, this model provides better protection against congestion situations, thus being an attractive solution in multi-tenant environments.

The throughput-based specification is challenging for the infrastructure provider since a certain throughput should be guaranteed regardless of the underlying resources. The management system can continuously monitor the network throughput to detect a lack of capacity and provide the required changes in the network infrastructure. This problem, which is out of the scope of this work, is addressed in [45], where a self-planning entity runs an iterative algorithm to derive planning actions such as adding or relocating SCs in order to meet the capacity needs. On the other hand, the resource-based options require simpler cell planning, since there is a direct mapping between the number of RBs and the number of SCs. However, the existence of different levels of isolation entails a complex scenario with a greater variety of resource allocation strategies. Each strategy represents a particular degree of flexibility for allocating spectrum chunks. The infrastructure provider can leverage this flexibility to apply its own resource allocation policies. Example policies are minimizing inter-cell interference or maximizing slice isolation. The latter can facilitate the application of interference-mitigating strategies separately for each slice.

Since the throughput-based specification is analyzed in [45], this work focuses on the different resource-based specification approaches, analyzing their impact on isolation, performance and flexibility for enforcing operator's policies. With respect to other conditioning parameters, for the sake of simplicity, in this work the requested capacity is expressed as a guaranteed downlink capacity. In addition, the per-UE granularity level is excluded from the analysis since the requested per-UE capacity could be expressed on a per-cell or network basis by simply considering the maximum number of sessions or UEs.

B. RESOURCE-BASED SPECTRUM SHARING STRATEGIES

Considering the capacity reservation type, granularity level, and isolation level parameters, there are six representative resource-based capacity specifications that are mapped to different strategies of spectrum sharing between slices. These strategies determine an average resource allocation at the planning phase. Thus, they are compatible with elastic resource allocation performed during the operation phase [33].

1) STRATEGY 1 (S1): PER-CELL RESOURCE-BASED CAPACITY PLANNING WITH INTER-CELL ISOLATION

In this strategy, the requested value of spectrum chunks, $v^{(s)}$, is a constant value per cell. In this way, all the SCs should allocate the same number of chunks for the slice. Furthermore, due to the inter-cell isolation requirement, other slices cannot use the selected chunks. From the infrastructure provider' interests, concentrating the required resources over the network into a minimum number of chunks is an attractive solution to leave room for other slices demanding inter-cell isolation. This represents the most likely situation in a multi-tenant scenario. For example, the provider can simply select a number of chunks per slice from the set R. However, in case there is a plethora of available chunks (i.e. not allocated to other slices), the resource allocation could target the reduction of inter-cell interference in the network by selecting disjoint sets of chunks among SCs.

This strategy of resource allocation does not provide flexibility to allocate resources due to the hard constraint on the number of chunks per cell. Consequently, S1 will not be able to adapt to spatial variations of the traffic demand. On the contrary, the high isolation level enables good protection against inter-slice interference and congestion situations.

2) STRATEGY 2 (S2): PER-NETWORK RESOURCE-BASED CAPACITY PLANNING WITH INTER-CELL ISOLATION

The requested value of spectrum chunks in S2 is defined on a network basis. It represents a soft constraint since a number $v^{(s)}$ of chunks in the network is to be freely distributed among the SCs. This problem is equivalent to targeting a specific per-cell average of the number of chunks. S2 differs from S1 because the former allows some variations in the number of chunks per SC in order to cope with spatial traffic variations.

The S2 is divided into three steps:

1) Select a set $R^{(s)}$ of chunks for the slice (e.g. chunk #1 and #2). This number of chunks should be high enough to allocate $v^{(s)}$ chunks among the SCs. As in S1, the provider typically concentrates the required resources into a minimum number of chunks in the network.
2) Determine the number of chunks $|R_{b,s}|$ to allocate at each SC considering the spatial traffic variations and the requested value $v^{(s)}$. This number is chosen to be proportional to the estimated value of the slice's service demand in the cell area, whose actual value is defined as:

$$D^{(b,s)} = \sum_{\substack{u|b=\Gamma(u)\\ b\in B_s, u\in U_s}} D_u, \qquad (10)$$

where D_u is the service demand of the UE u and $\Gamma(u)$ returns the serving SC. The estimated value, $\hat{D}^{(b,s)}$, is based on the method proposed in [45], considering the service demand D_u and the correlation that can be expected between the slice's service demand and the actual network's service demand. To ensure accessibility from any location, at least one chunk is allocated per SC. In addition, the total number of allocated chunks should be equal to the required value $v^{(s)}$, i.e.:

$$v^{(s)} = \sum_{b\in B_s} \left|R_{b,s}\right|. \tag{11}$$

Algorithm 1 describes the procedure to determine the number of chunks per SC, where $\widehat{r}_{b,s}$ stands for the actual number of chunks and $r'_{b,s}$ is the targeted number of chunks, calculated as:

$$r'_{b,s} = \frac{\hat{D}^{(b,s)}}{K}, \tag{12}$$

where K is the constant of proportionality between the cell demands and the number of chunks. Specifically, the chunks are allocated following an iterative process where only one chunk is allocated for each SC within the same iteration. If, at a certain iteration in the loop starting at the line 4, the number of allocated chunks in a SC reaches the targeted value $r'_{b,s}$ calculated at the line 3, no more chunks will be allocated in that SC at this stage. However, if there are still chunks to allocate to the slice after the loop, the process continues allocating chunks consecutively in each SC until the stopping condition at the line 12 is satisfied.

Algorithm 1 Calculation of the Number of Chunks for SCs

```
1:  Inputs: $\hat{D}^{(b,s)}$, $B_s$, $v^{(s)}$
2:  Initialize $\widehat{r}_{b,s} = 0$; $r'_{b,s} = 0$
3:  Compute $r'_{b,s}$, $b \in B_s$
4:  while $\sum_{b\in B_s} \widehat{r}_{b,s} < v^{(s)}$ and $\sum_{b\in B_s} r'_{b,s} > 0$ do
5:      for $b \in B_s$
6:          if $r'_{b,s} > 0$ and $\sum_{b\in B_s} \widehat{r}_{b,s} < v^{(s)}$ then
7:              $\widehat{r}_{b,s} = \widehat{r}_{b,s} + 1$;
8:              $r'_{b,s} = r'_{b,s} - 1$;
9:          end if
10:     end for
11: end while
12: while $\sum_{b\in B_s} \widehat{r}_{b,s} < v^{(s)}$ do
13:     for $b \in B_s$
14:         if $\sum_{b\in B_s} \widehat{r}_{b,s} < v^{(s)}$ then $\widehat{r}_{b,s} = \widehat{r}_{b,s} + 1$; then
15:         end if
16:     end for
17: end while
18: return $\widehat{r}_{b,s}$
```

1) Allocate a set $R_{b,s}$ of chunks to every SC b given the set $R^{(s)}$ from which they are selected and the required number of chunks per SC, $\widehat{r}_{b,s}$. The selection is made according to the algorithm proposed in [45], which minimizes inter-cell interference. In this algorithm, the chunk allocation in a SC is performed so that the SC-to-SC distance between the given SC and the closest neighboring SC using the same chunk is the maximum possible. After the execution of the algorithm, the cardinality of the set $R_{b,s}$ for each SC should be equal to $\widehat{r}_{b,s}$.

Applying S2 to a given slice leads to a more efficient resource usage than S1 since it fits better the spatial demands. However, in global terms, this efficiency could be small if other slices that are more loaded cannot share RBs with this slice. Such an effect is consequence of the high isolation level of S2, which also occurs in S1.

3) STRATEGY 3 (S3): PER-CELL RESOURCE-BASED CAPACITY PLANNING WITH INTRA-CELL ISOLATION

In S3, the per-cell definition of the requested value $v^{(s)}$ means that the number of chunks is the same for all the SCs. Consequently, this strategy is not adequate, like S1, to fit the spatial traffic variations. Unlike the previous strategies, the intra-cell isolation requirement in S3 enables that other slices can use the same chunk in a different SC, thus ensuring isolation between slices only within the area of a SC.

The chunk allocation can be targeted to minimize co-channel interference by following the algorithm proposed in [45] or, alternatively, it can be oriented to maximize isolation by reusing the same chunks across the SCs if they are available. The former approach is assumed in this work since it has a better impact on the network performance. Consequently, the algorithm proposed in [45] is applied considering that, at each SC, the set of chunks belonging to other slices are not eligible. While in S2 the set of candidate chunks (given by $R^{(s)}$) is the same for all SCs, in S3 it may differ from SC to SC. Accordingly, $R^{(b,s)}$ represents the candidate chunks at each SC b for the slice s.

This strategy gives providers more flexibility to distribute RBs among slices since the required isolation level is lower. However, it is limited by the requirement of a fixed number of chunks per cell, which may lead to a suboptimal matching between demand and resources.

4) STRATEGY 4 (S4): PER-NETWORK RESOURCE-BASED CAPACITY PLANNING WITH INTRA-CELL ISOLATION

In this case, the per-network definition of the requested value $v^{(s)}$ provides flexibility to adapt to the spatial traffic variations, i.e. the number of chunks per SC can vary to meet the particular traffic demand at each cell. The only constraints are that each SC has at least one chunk allocated and that the total number of chunks allocated in the network (or cluster) is equal to $v^{(s)}$, i.e. the condition in (11) is satisfied. Since isolation is only required within the cell area, the provider has even greater flexibility than S2 to perform the chunk allocation. Based on this, the process is composed of the following steps:

1) Determine the number of chunks $|R_{b,s}|$ to allocate at each SC based on the spatial demand distribution and the requested value $v^{(s)}$. This step is the same as the step 2 in S2.
2) Allocate a set $R_{b,s}$ of chunks to every SC b given the set $R^{(b,s)}$ from which they are selected and the required number of chunks per SC, $\widehat{r}_{b,s}$, obtained in the step 1. To minimize the effect of inter-cell interference, the algorithm proposed in [45] is applied. It may happen that $|R^{(b,s)}|$ is lower than $\widehat{r}_{b,s}$, meaning that the number of candidate chunks is not enough to reach the targeted value in a certain SC. In this case, the provider has enough flexibility to allocate the missing chunk(s) in a different SC without affecting the requested (per-network) value $v^{(s)}$.

A major advantage of this strategy is the greater flexibility to allocate RBs to slices with varying demands in a successful way. The gain will be larger as the number of slices planned with this strategy increases.

5) STRATEGY 5 (S5): PER-CELL RESOURCE-BASED CAPACITY PLANNING WITH NO ISOLATION

This strategy establishes a specific number of chunks per SC. Therefore, the number of chunks cannot be adapted to the cell demand at each SC. The main difference with the previous strategies is that, in this case, resource isolation is not mandatory. The providers can exploit this idea to leave room for other slices demanding resource isolation. In particular, those slices that do not require isolation will share RBs within the same SC. The selection of chunks is based on the algorithm proposed in [45] in order to minimize co-channel interference. The algorithm takes as input the set $R^{(b,s)}$ of candidate chunks at each SC. Among these chunks, the algorithm prioritizes the ones shared with other slices. However, it is necessary to evaluate whether the SC is able to support or not the estimated $\hat{D}^{(b,s)}$. If not, the chunk is discarded from the candidate set.

Compared to the previous strategies, the S5 provides better resource usage as the traffic from different slices is aggregated into the same chunks. However, it is not optimal because the cell resources cannot be adapted to the spatial traffic variations. From the tenant's perspective, this strategy is suitable for services with no stringent requirements in terms of capacity or latency. In these cases, the tenant can achieve significant cost savings at the expense of greater uncertainty in performance due to the traffic load variations of other slices.

6) STRATEGY 6 (S6): PER-NETWORK RESOURCE-BASED CAPACITY PLANNING WITH NO ISOLATION

The requested value in S6 is defined on a per-network basis. It means that the number of chunks at each SC can be adjusted to meet its traffic needs while guaranteeing a total number of chunks, $v^{(s)}$, allocated in the network. In addition, this strategy does not require resource isolation, which gives the provider the possibility to allocate the chunks being shared with other slices.

The process behind S6 is similar to the strategies S2 and S4, which also define a network-wide capacity value. First, the number of chunks to allocate at each SC is calculated based on the spatial demand distribution. Second, a set of chunks is selected from a set of available chunks based on the algorithm proposed in [45], giving priority to the shared ones without causing overload. Note that, if the traffic of two slices is assumed highly correlated, enforcing S6 is similar to applying it to a single slice carrying the aggregated traffic of both slices.

This strategy provides more efficient resource usage than S5 since the number of chunks at each SC can be fitted to the spatial traffic conditions. However, since no isolation is allowed, it retains the same disadvantages with regard to the impact of the traffic load variations of other slices.

V. PERFORMANCE EVALUATION

A. SIMULATION SCENARIO

In order to evaluate the proposed spectrum planning strategies, numerical examples and simulation results are presented in this section. The service area is 1.5 km × 1.5 km. It covers an urban environment with a set of deployed SCs. More specifically, the deployment scenario comprises the following: (i) for each slice, the statistical characterization of the traffic demand, which is non-uniformly distributed over the considered area and spatially cross-correlated with other slices; and (ii) a set of SCs deployed in the scenario according to the spatial variations of the aggregated traffic demand. The deployment scenario is simulated following a snapshot-based model, where each snapshot represents a random realization of the demand distribution. The different realizations of the same traffic probability distribution (i.e. varying the positions of the UEs) ensure reliable statistical significance analysis. The deployment scenario for 95% correlated demand between the slices A and B is shown in Fig. 3, where the triangles represent the location of the deployed SCs and the colored contour lines indicate the aggregated traffic demand density. As observed, the areas with higher traffic densities are provided with more SCs that are strategically located to serve the demand maximizing resource usage efficiency. The crosses in the figure represent the UE locations for a certain realization of the traffic probability distributions. The color of the crosses indicates the slice (A or B) to which the UE is connected. Finally, Table 1 summarizes the main parameters of the simulations. The requested value $v^{(s)}$ is calculated as a function of the considered RB occupancy in the network.

One issue regarding the implementation of the network model is that, due to the mutual interference between SCs, there is a dependency relation between the cell loads [54]. To reduce the computational complexity of the cell load

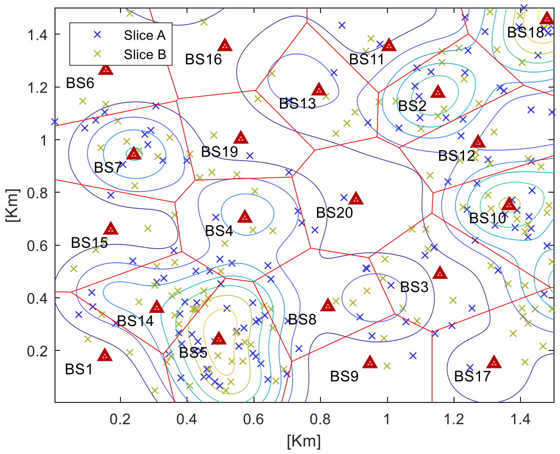


FIGURE 3. SC deployment with non-uniform traffic demand distribution and 95% correlated between slices.

TABLE 1. Simulation parameters.

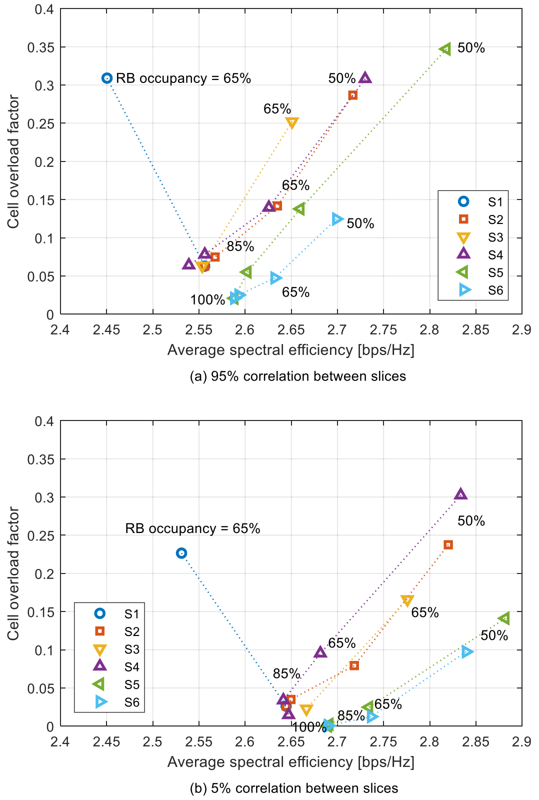
Parameter	Configuration
Cellular environment	Urban, 1.5km x 1.5km
Number of SCs	20
Operating frequency	5 GHz
5G numerology (μ)	0
System bandwidth	120 MHz
Minimum allocation unit size (chunk)	20 MHz (106 RBs)
Propagation (path loss, shadowing)	UMi model [52]
SC antenna directivity	omni-directional
SC antenna height	6 m
UE antenna height	1.5 m
SC antenna gain	2 dBi
UE thermal noise	-174 dBm/Hz
UE noise figure	9 dB
Target SINR at cell-edge	9 dB [53]
UE minimum SINR	-10 dB [50]
UE maximum SINR	30 dB [50]
SC TX power range	[25-33] dBm
Number of UEs	250
UE service demand	5 Mbps
Resource scheduling scheme	Round-Robin
Number of slices	2
Proportion of UEs per slice	[50, 50] %
Traffic correlation between slices	95%, 5%
RB occupancy in the network	[50-100] %
Number of demand realizations	100

factor [see (4) and (5)], the per-UE spectral efficiency SE_u is substituted by a cell-specific average value, which is taken from previous evaluations (i.e. from other snapshots). There is also a dependency relation between the $SINR(u,r)$ and $\Gamma(u)$, since the latter is based on the SINR to determine the serving SC. To simplify the procedure, the $SINR(u,r)$ is first estimated using $\Gamma(u)$ based on the strongest received power, which is calculated in (3). Then, the obtained values are used to compute $\Gamma(u)$ based on the SINR and, lastly, the $SINR(u,r)$.

B. PERFORMANCE ANALYSIS OF THE SPECTRUM PLANNING STRATEGIES

The first experiment provides a comparison between the different planning strategies regarding both network and service performance aspects. Specifically, the network performance is assessed using the cell overload factor and spectral efficiency metrics as defined in (6) and (7), respectively. The service performance is evaluated through the unsatisfied UE rate, which is defined as the fraction of UEs experiencing a throughput below the 25% of the UE service demand. The study is performed for two levels of correlation, 95% and 5%, between the traffic demands of the slices in the spatial domain. The high correlation value may represent slices that provide different services from the same tenant or the same service from different tenants. The low correlation value is more likely for slices owned by different tenants providing disparate services. To ensure a fair comparison, the strategies are evaluated for the same percentages of allocated spectrum chunks in the network. The RB occupancy affects network performance in the sense that, with high occupancy, the network will have more capacity, thus reducing the cell overload factor for the existing slices; however, it also reduces the possibilities to accept new slices, especially the ones that require resource isolation. The considered RB occupancy levels in the simulations are 50, 65, 85 and 100%, except for S1 and S3 for which the 50 and 85% values are not feasible. The latter is due to the simultaneous occurrence of the following factors: the proportion of UEs (and chunks) per slice is 50%; the specified number of chunks per cell must be an integer; and the 50% and 85% of the number of chunks in the system bandwidth (i.e. three and five chunks, respectively) is not divisible by the number of slices.

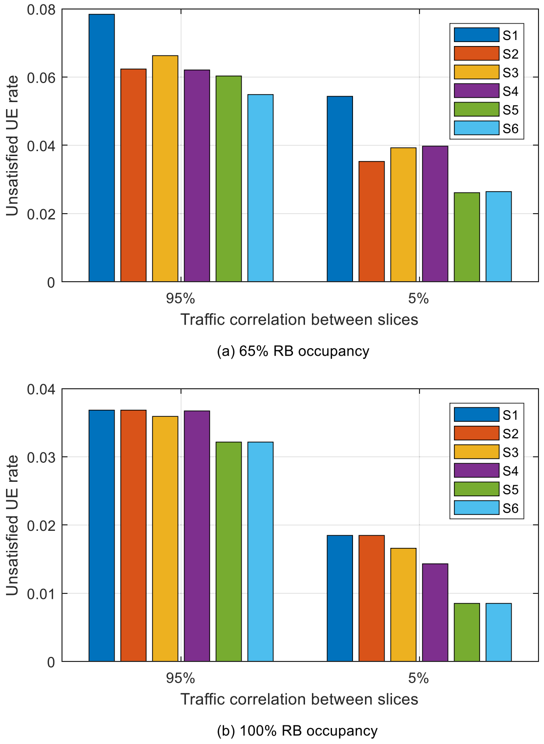
The network performance metrics are represented against each other in Fig. 4(a-b) for the two correlation levels. The dotted lines connect the performance values for the different RB occupancy levels following a sequential order (50-65-85-100% for S2, S4, S5, S6 and 65-100% for S1, S3). The two scenarios are evaluated with the same number of UEs; however, the values of the metrics are better with a lower correlation because the aggregated load of the two slices is more regularly distributed in the scenario. As observed, the trade-off between the cell overload factor and the average spectral efficiency is applicable to all the strategies except for S1 due to its poor matching between traffic demand and network resources. In fact, such inefficiency is reflected in Fig. 4 by the extremely high values of the cell overload factor for a 65% of RB occupancy. The S3 also results in a high cell overload factor since the number of allocated chunks per cell cannot be adapted to the cell load. However, it provides a better average spectral efficiency than S1 because the S3 utilizes the entire system bandwidth to allocate the 65% of chunks in the network, while the S1 only uses a fraction of the system bandwidth with reuse-1. The strategies with no isolation, S5 and S6, provide the best trade-off between the two network metrics as their dotted lines in the figure are closer to the bottom right corner, especially for the case of a low demand correlation. This is a reasonable result since resource sharing typically leads to optimal network performance. However, as discussed later, the increased performance is at the expense of no isolation between slices. Behind these strategies, the



(a) 95% correlation between slices

(b) 5% correlation between slices

FIGURE 4. **Evaluation of network metrics for different percentages of chunk allocation and correlation levels between slices.**



(a) 65% RB occupancy

(b) 100% RB occupancy

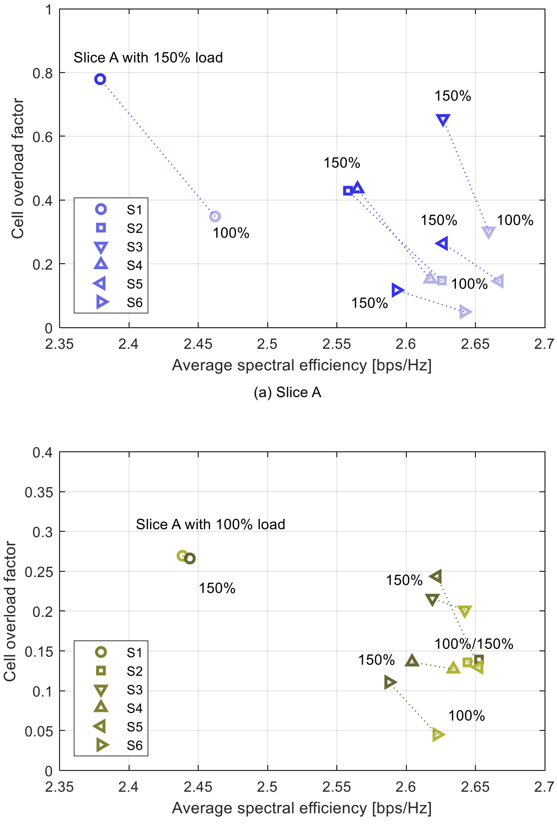
FIGURE 5. **Evaluation of the unsatisfied UE rate for different percentages of chunk allocation and correlation levels between slices.**

S2 and S4 show good network performance while at the same time they provide resource isolation. Since the required capacity in these strategies is defined on a per-network basis, they enable some adaptation to the spatial traffic variations. Therefore, the resource usage is more efficient than in S1 and S3, where the slice specification entails a more rigid chunk allocation in the SCs. Such flexibility in allocating chunks is especially useful when the traffic demand of the slices is poorly correlated. Specifically, a slice can benefit from allocating additional chunks in overloaded areas where other slices are unloaded. The S2 and S4 achieve similar performance for the case of 95% correlated traffic, as shown in Fig. 4(a). However, in the case of 5% correlated traffic [see Fig. 4(b)] and high RB occupancy (above 85%), the S4 overcomes the performance of S2. In particular, for a 100% of RB occupancy, the S4 obtains a cell overload factor of 1.4%, while the S2 obtains 2.6%. Thus, the greater flexibility offered by the S4 due to the intra-cell isolation (as opposed to the inter-cell isolation of S2) entails an increased performance only if the number of allocated chunks in the network is sufficiently high.

Regarding service performance, Fig. 5(a-b) shows the mean of the unsatisfied UE rate obtained by each planning strategy for two correlation levels of the traffic demand, 95% and 5%, and two levels of RB occupancy, 65% and 100%. With 65% of RB occupancy, it is observed that the S1 results in the worst service performance level. However, with a 100% of RB occupancy [see Fig. 5(b)], the fraction of spectrum used by the S1 matches the system bandwidth. Consequently, its unsatisfied UE rate is similar to other strategies. The S2, S3 and S4 provide similar unsatisfied UE rate, being better than S1 but worse than S5 and S6. With a 100% of RB occupancy and 5% of correlated traffic, the S4 provides better performance than S2 and S3 because, in this situation, the network further benefits from a more flexible chunk allocation. Lastly, the S5 and S6 result in the best performance level as they share chunks between slices. This gain is more pronounced with a 5% correlated traffic, reducing to the half the unsatisfied UE rate obtained by the S1.

C. ANALYSIS OF THE IMPACT ON RESOURCE ISOLATION

The following experiment evaluates the sensitivity of the performance metrics to the load variations of the slices. The scenario assumes a 95% of correlated traffic between slices and a 65% of RB occupancy. Initially, the network load is given by the configuration in Table 1, i.e. 250 UEs equally distributed between slices. After the initial stage, the load of the slice *A* increases by 50% (reaching 187 UEs), maintaining the same spatial distribution. The load of the slice *B* is not modified. Fig. 6(a-b) shows the performance comparison between both situations concerning network metrics. The results are given per slice in order to highlight the impact of the increased load in the slice *A* on each slice separately. As observed in Fig. 6(a), the marks indicating the performance level with a 100% of load are shifted towards the upper left corner



(a) Slice A

(b) Slice B

FIGURE 6. Evaluation of network metrics when the slice A increases its traffic demand by 50%. A 95% of correlated traffic between slices and a 65% of RB occupancy in the network are assumed.

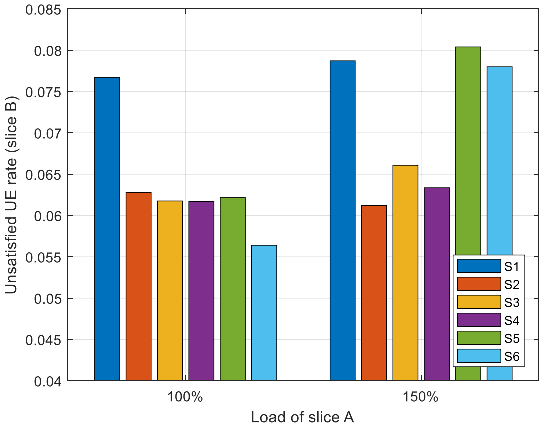


FIGURE 7. Evaluation of the unsatisfied UE rate for the slice B before and after the slice A increases its traffic demand by 50%.

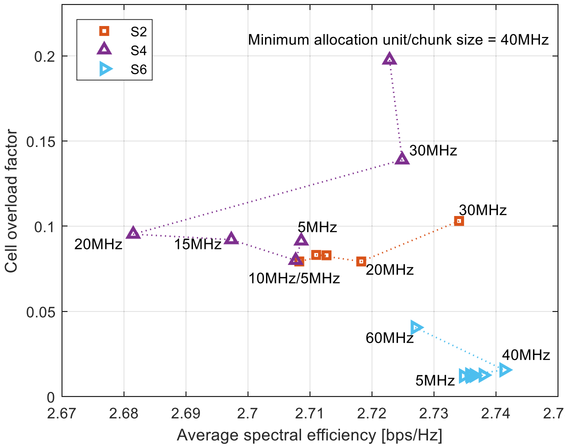


FIGURE 8. Evaluation of the impact of modifying the minimum allocation unit size for different planning strategies. A 5% of correlated traffic between slices and a 65% of RB occupancy in the network are assumed.

of the figure when such a load increases by 50%, meaning degradation of both metrics. Roughly, the cell overload factor is doubled in all the strategies except for the S2 and S4, where this increase is even greater. In Fig. 6(b), the results for the slice *B* depend on the degree of isolation enforced by the planning strategy. In the case of inter-cell isolation, the S1 and S2 maintain the same performance level before and after the load increase, thus providing full protection against the traffic variations of other slices. On the contrary, the other strategies are impacted to an extent that depends on the isolation level. The S3 and S4 result in a slight degradation since they perform intra-cell isolation. The S5 and S6 lead to a significantly worse performance, particularly, in terms of the cell overload factor, whose value is approximately doubled. As these strategies do not perform resource isolation, the impact of traffic variations from other slices is the greatest possible.

Fig. 7 represents the unsatisfied UE rate for the slice *B* before and after the load of the slice *A* increases by 50%. With a 100% of load, the results are in line with Fig. 6(b). Specifically, the S1 provides the worst service performance, while the S6 is slightly better than the other strategies. However, with a 150% of load, the degradation in the S5 and S6 leads to an unsatisfied UE rate that is similar to the performance of the S1, while the other strategies are hardly affected.

D. ANALYSIS OF THE SCALABILITY OF THE STRATEGIES

In this experiment, the impact of varying the chunk size is analyzed. To avoid a larger computational load, the system bandwidth is not modified, maintaining the same value as in Table 1, i.e. 120 MHz. In return, the minimum allocation unit size (i.e. the chunk size) is modified by sweeping the parameter through the following values: 60, 40, 30, 20, 15, 10 and 5 MHz, which are equivalent to 2, 3, 4, 6, 8, 12 and 24 chunks in the network. Each chunk comprises 324, 216, 160, 106, 79, 52 and 25 RBs, respectively. For example, the previous experiments are carried out with 20 MHz of minimum allocation unit size, which means 6 chunks for the system bandwidth and 100 RBs per chunk. Depending on the planning strategy, some values of the minimum allocation unit size may not be feasible given the configuration in Table 1. In addition, the evaluated cases assume a 5% of correlated traffic between slices and a 65% of RB occupancy.

The results are shown in Fig. 8 with focus on the strategies for which the slice capacity is specified on a per-network basis. The dotted lines connect the performance values for the different minimum allocation unit sizes following a sequential order (5-10-15-20-30 MHz for S2, 5-10-15-20-30-40 MHz for S4 and 5-10-15-20-30-40-60 MHz for S6). Such strategies enable flexible chunk allocation, so that they can further benefit from increasing the number of chunks. It is observed that, as the number of chunks increases (i.e. the minimum allocation unit size decreases), the strategies saturate beyond a certain value. In particular, the cell overload factor saturates when the minimum allocation unit size is below 20 MHz. This value, used in the previous experiments, is applicable to the three planning strategies, as shown in Fig. 8. Since there is no substantial gain for chunk sizes below 20 MHz, this limit provides a good trade-off between performance and complexity. Lastly, it is also observed that the S4 is more sensitive to the variations of the chunk size than the other strategies.

VI. CONCLUSION

This work has addressed the problem of spectrum planning for a 5G sliced network. To facilitate the transition to the new 5G paradigm, a business-driven model has been proposed to define tenant's requirements from a perspective of network resources. Then, following this model, different spectrum planning strategies for RAN slicing have been developed based on various levels of resource isolation and granularity. These strategies focus on the resource-oriented (as opposed to throughput-oriented) capacity specification. Each possible strategy gives the infrastructure provider different degrees of flexibility to allocate resources. By leveraging this flexibility, the provider can efficiently adapt the network resources to the traffic demands of the slices. The proposed spectrum planning strategies have been evaluated in a 5G sliced network of SCs through snapshot-based simulations. The results show that the strategies with no isolation provide the best network performance due to a more efficient resource usage. The strategies based on a per-network capacity specification, as opposed to a per-cell definition, enable better adaptation to the spatial traffic variations, resulting in higher performance for low network resource occupancy and high traffic correlation between slices. The strategies with intra-cell or inter-cell isolation provide similar protection against inter-slice interference. However, for high resource occupancy and low traffic correlation, the intra-cell isolation results in better performance because of the greater flexibility for adapting resources to the slices' demands.

An interesting direction for future work extensions is the reallocation of resources in the network when a new slice request arrives. The new slice, based on its capacity specifications, may require some changes on the current resource allocation for its successful deployment. Moreover, when the slice becomes operative, the resources could also be dynamically allocated to cope with temporal variations of the traffic demands. AI techniques can also be applied to automate and optimize these tasks.

REFERENCES

[1] R. N. Mitra and D. P. Agrawal, "5G mobile technology: A survey," *ICT Express*, vol. 1, no. 3, pp. 132–137, Dec. 2015.

[2] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[3] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis, "Multi-tenant slicing for spectrum management on the road to 5G," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 118–125, Oct. 2017.

[4] S. M. A. Kazmi, L. U. Khan, N. H. Tran, and C. S. Hong, *Network Slicing for 5G and Beyond Networks*. Cham, Switzerland: Springer, 2019.

[5] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.

[6] Y. Zhang, *Network Function Virtualization: Concepts and Applicability in 5G Networks*. Hoboken, NJ, USA: Wiley, 2018.

[7] H.-T. Chien, Y.-D. Lin, C.-L. Lai, and C.-T. Wang, "End-to-End slicing as a service with computing and communication resource allocation for multi-tenant 5G systems," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 104–112, Oct. 2019.

[8] R. Su, D. Zhang, R. Venkatesan, Z. Gong, C. Li, F. Ding, F. Jiang, and Z. Zhu, "Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models," *IEEE Netw.*, vol. 33, no. 6, pp. 172–179, Nov. 2019.

[9] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[10] *Feasibility Study on New Services and Markets Technology Enablers; Stage 1, Version 14.2.0*, 3GPP, document TR 22.891, Sep. 2016.

[11] *NR; NR and NG-RAN Overall Description; Stage 2, Version 16.0.0*, 3GPP, document TR 38.300, Dec. 2019.

[12] O. Bulakci and E. Pateromichelakis, "Slice-aware 5G dynamic small cells," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakesh, Morocco, Apr. 2019, pp. 1–6.

[13] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[14] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984.

[15] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 80–87, May 2017.

[16] D. A. Chekired, M. A. Togou, L. Khoukhi, and A. Ksentini, "5G-Slicing-Enabled scalable SDN core network: Toward an ultra-low latency of autonomous driving service," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1769–1782, Aug. 2019.

[17] L. Ma, X. Wen, L. Wang, Z. Lu, and R. Knopp, "An SDN/NFV based framework for management and deployment of service based 5G core network," *China Commun.*, vol. 15, no. 10, pp. 86–98, Oct. 2018.

[18] J. Ordonez-Lucena, O. Adamuz-Hinojosa, P. Ameigeiras, P. Munoz, J. J. Ramos-Munoz, J. F. Chavarria, and D. Lopez, "The creation phase in network slicing: From a service order to an operative network slice," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Ljubljana, Slovenia, Jun. 2018, pp. 1–36.

[19] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, and T. Braun, "Network slices toward 5G communications: Slicing the LTE network," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 146–154, Aug. 2017.

[20] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

[21] P. L. Vo, M. N. H. Nguyen, T. A. Le, and N. H. Tran, "Slicing the edge: Resource allocation for RAN network slicing," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 970–973, Dec. 2018.

[22] J. Garcia-Morales, M. C. Lucas-Estan, and J. Gozalvez, "Latency-sensitive 5G RAN slicing for industry 4.0," *IEEE Access*, vol. 7, pp. 143139–143159, 2019.

[23] J. J. Escudero-Garzas, C. Bousono-Calzon, and A. Garcia, "On the feasibility of 5G slice resource allocation with spectral efficiency: A probabilistic characterization," *IEEE Access*, vol. 7, pp. 151948–151961, 2019.

[24] 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 Conf. Comput. Commun. (INFOCOM)*, Paris, France, Apr. 2019, pp. 442–450.

[25] Z. Kotulski, T. W. Nowak, M. Sepczuk, and M. A. Tunia, "5G networks: Types of isolation and their parameters in RAN and CN slices," *Comput. Netw.*, vol. 171, Apr. 2020, Art. no. 107135.

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

[27] T. Guo and A. Suarez, "Enabling 5G RAN slicing with EDF slice scheduling," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2865–2877, Mar. 2019.

[28] D. Marabissi and R. Fantacci, "Highly flexible RAN slicing approach to manage isolation, priority, efficiency," *IEEE Access*, vol. 7, pp. 97130–97142, 2019.

[29] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[30] S. O. Oladejo and O. E. Falowo, "5G network slicing: A multi-tenancy scenario," in *Proc. Global Wireless Summit (GWS)*, Cape Town, South Africa, Oct. 2017, pp. 88–92.

[31] S. N. Khan, L. Goratti, R. Riggio, and S. Hasan, "On active, fine-grained RAN and spectrum sharing in multi-tenant 5G networks," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.

[32] O. Al-Khatib, W. Hardjawana, and B. Vucetic, "Spectrum sharing in multi-tenant 5G cellular networks: Modeling and planning," *IEEE Access*, vol. 7, pp. 1602–1616, 2019.

[33] D. M. Gutierrez-Estevez, Y. Wang, M. Gramaglia, A. D. Domenico, G. Dandachi, S. Khatibi, D. Tsolkas, I. Balan, A. Garcia-Saavedra, and U. Elzur, "Artificial intelligence for elastic management and orchestration of 5G networks," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 134–141, Oct. 2019.

[34] V. Sciancalepore, X. Costa-Perez, and A. Banchs, "RL-NSB: Reinforcement learning-based 5G network slice broker," *IEEE/ACM Trans. Netw.*, vol. 27, no. 4, pp. 1543–1557, Aug. 2019.

[35] H. Xiang, S. Yan, and M. Peng, "A realization of fog-RAN slicing via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2515–2527, Apr. 2020.

[36] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, "Dynamic reservation and deep reinforcement learning based autonomous resource slicing for virtualized radio access networks," *IEEE Access*, vol. 7, pp. 45758–45772, 2019.

[37] X. Chen, Z. Zhao, C. Wu, M. Bennis, H. Liu, Y. Ji, and H. Zhang, "Multi-tenant cross-slice resource orchestration: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2377–2392, Oct. 2019.

[38] R. Li, Z. Zhao, Q. Sun, C.-L. I, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.

[39] A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, "DeepSlice: A deep learning approach towards an efficient and reliable network slicing in 5G networks," in *Proc. IEEE 10th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, New York, NY, USA, Oct. 2019, pp. 762–767.

[40] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 361–376, Feb. 2020.

[41] 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. 33137–33147, 2018.

[42] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, May 2018.

[43] J. Perez-Romero, O. Sallent, R. Ferrus, and R. Agusti, "On the configuration of radio resource management in a sliced RAN," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Taipei, Taiwan, Apr. 2018, pp. 1–6.

[44] J. Gang and V. Friderikos, "Inter-tenant resource sharing and power allocation in 5G virtual networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7931–7943, Aug. 2019.

[45] P. Munoz, O. Sallent, and J. Perez-Romero, "Self-dimensioning and planning of small cell capacity in multitenant 5G networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4552–4564, May 2018.

[46] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On the automation of RAN slicing provisioning and cell planning in NG-RAN," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Ljubljana, Slovenia, Jun. 2018, pp. 37–42.

[47] *NR; User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone, Version 16.2.0*, 3GPP, document TS 38.101-1, Dec. 2019.

[48] *NR; Physical Channels and Modulation, Version 16.0.0*, 3GPP, document TS 38.211, Dec. 2019.

[49] *NR; Physical Layer Procedures for Data, Version 16.0.0*, 3GPP, document TS 38.214, Dec. 2019.

[50] *Study on New Radio Access Technology: Radio Frequency (RF) and co-Existence Aspects, Version 14.2.0*, 3GPP, document TR 38.803, Sep. 2017.

[51] *Generic Network Slice Template, Version 2.0*, document NG.116, GSM Association, London, U.K., Oct. 2019.

[52] *Study on Channel Model for Frequencies From 0.5 to 100 GHz, Version 16.1.0*, 3GPP, document TR 38.901, Dec. 2019.

[53] D. Lopez-Perez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, 4th Quart., 2015.

[54] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.

PABLO MUÑOZ received the M.Sc. and Ph.D. degrees in telecommunication engineering from the University of Málaga, Málaga, Spain, in 2008 and 2013, respectively. He is currently an Assistant Professor with the Department of Signal Theory, Telematics, and Communications, University of Granada, Granada, Spain. He has published more than 50 articles in peer-reviewed journals and conferences. He is the coauthor of four international patents. His research interests include radio access network planning and management, application of artificial intelligence tools in radio resource management, and virtualization of wireless networks.

ÓSCAR ADAMUZ-HINOJOSA received the B.Sc. and M.Sc. degrees in telecommunications engineering from the University of Granada, Spain, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Department of Signal Theory, Telematics and Communication. His research interests are focused on SDN, NFV, and network slicing in 5G systems.

JORGE NAVARRO-ORTIZ received the M.Sc.E.E. degree from the University of Malaga, Spain, in 2001, and the Ph.D. degree from the University of Granada, Spain. Then, he worked at Nokia Networks, Optimi/Ericsson, and Siemens. He started working as an Assistant Professor at the University of Granada, in 2006. He is currently an Associate Professor with the Department of Signal Theory, Telematics and Communications, University of Granada. His research interests include wireless technologies for the IoT, such as LoRaWAN and 5G.

ORIOL SALLENT is currently a Professor with the Universitat Politècnica de Catalunya (UPC). He has participated in a wide range of European projects with diverse responsibilities as a Workpackage Leader and a Coordinator Partner and contributed to standardization bodies, such as 3GPP, IEEE, and ETSI. He has published over 200 articles mostly in IEEE journals and conferences. His research interests include cognitive management in cognitive radio networks, self-organizing networks, radio network optimization, and QoS provisioning in heterogeneous wireless networks.

JORDI PÉREZ-ROMERO (Member, IEEE) is currently a Professor with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. He has been involved in different European projects and in projects for private companies. He has published more than 200 articles in international journals and conferences. He has been working in the field of wireless communication systems, with particular focus on radio resource management, cognitive radio networks, and network optimization.