 DOI 10.1109/MWC.2017.1600220WC

On Radio Access Network Slicing from a Radio Resource Management Perspective

O. Sallent, J. Pérez-Romero, R. Ferrús, R. Agustí

Universitat Politècnica de Catalunya (UPC)

[sallent, jorperez, ferrus, ramon]@tsc.upc.edu

***Abstract*—Network slicing is a fundamental capability for future Fifth Generation (5G) networks to properly support current and new envisaged application scenarios. Network slicing facilitates a cost-effective deployment and operation of multiple logical networks over a common physical network infrastructure such that each network is customized to best serve the needs of specific applications (e.g., mobile broadband, Internet of Things applications) and/or communications service providers (e.g., special purpose service providers for different sectors such as public safety, utilities, smart city, and automobiles). Slicing a Radio Access Network (RAN) becomes particularly challenging due to the inherently shared nature of the radio channel and the potential influence that any transmitter may have on any receiver. In this respect, this paper analyzes the RAN slicing problem in a multi-cell network in relation to the Radio Resource Management (RRM) functionalities that can be used as a support for splitting the radio resources among the RAN slices. Four different RAN slicing approaches are presented and compared from different perspectives, such as the granularity in the assignment of radio resources and the degrees of isolation and customization.**

***Keywords*—Network slicing; RAN slicing; Multi-tenancy; RAN sharing; 5G networks**

INTRODUCTION

The evolution of mobile communications systems towards the so-called Fifth Generation (5G) must face the challenging requirements of Mobile BroadBand (MBB) use cases as well as new diverse use cases associated with customers of new market segments and vertical industries (e.g., e-health, automotive, energy) [1]. Therefore, in addition to supporting the evolution of the current business models, 5G will expand to new ones. Partnerships will be established on multiple layers, ranging from sharing the infrastructure to exposing specific network capabilities as an end-to-end service or integrating partners' services into the 5G system through a rich and software-oriented capability set.

Given the large variety of requirements on network functionalities (in terms of, e.g., security and mobility), performance (e.g., ultra-low-latency and ultra-reliability) and associated business models, it is envisaged that the support of network slicing will be one of the pillars for building the 5G ecosystem.

Network slicing consists in exposing a common network infrastructure in the form of different customized logical networks, referred to as *network slices*. Each network slice can be tailored to support specific applications and/or be operated by a communications provider other than the owner of the physical network infrastructure. Examples of communications providers, referred herewith as *tenants*, are a mobile virtual network operator (MVNO) for the MBB consumer market or a vertical service provider. A network slice is composed of a collection of network functions and specific radio access technology settings that are combined for the specific use case or business model associated with a particular application and/or tenant [1].

Third Generation Partnership Project (3GPP) specifications have already added some support for sharing legacy network infrastructures among multiple tenants. Two main architectures are identified in [2], namely Multi-Operator Core Network (MOCN), in which a shared Radio Access Network (RAN) is directly connected to each of the multiple operators' core networks, and Gateway Core Network (GWCN), in which a shared core network is considered. Similarly, different works have addressed the virtualization of the RAN as a means to allow multiple operators to share the physical radio resources available at a base station [3][4][5]. Moreover, 3GPP has addressed initial architectural considerations about network slicing for 5G in [6], in which different service and operational requirements are identified. The network slicing concept is also present in the 5G architectures currently being developed by different research projects. In this respect, in [7], 5G-NORMA identifies different functional requirements for performing network slicing operations. In [8], METIS-II analyzes the implications of network slicing for the RAN architecture. In turn, SESAME [9] addresses multi-tenancy in Small Cell as a Service (SCaaS) and corresponding implications in terms of Self-Organizing Network (SON) features.

As stressed in [6], isolation among slices is a fundamental feature that ensures that the traffic of one slice does not negatively impact other slices. In addition, it is essential that this isolation is implemented in a way that leads to an efficient use of resources. Although mechanisms to ensure isolation are well established in multi-tenant data centers or wired network domains [10], isolation is currently one of the main open research challenges in wireless network virtualization [11]. When considering RAN slicing in a single-cell scenario, as in [3][4], isolation can be achieved by assigning an orthogonal set of physical radio resources for a certain period to each tenant in accordance with its requirements. Then, flexibility can be left to the tenant such that the radio resources are assigned to its customers during that period. Nevertheless, when considering the slicing of a multi-cell RAN, the concept of isolation is more challenging and deserves deeper analysis due to the potential interferences between transmitters of different tenants at different cells.

In this way, in addition to "traffic isolation", which refers to avoiding the situation in which a traffic overload generated by one tenant negatively affects the performance experienced by another tenant, "radio-

electrical isolation", which refers to avoiding mutual interferences at the air interface among the transmissions of different tenants, must also be considered. In this respect, this paper constitutes, to the authors' best knowledge, the first attempt to address the multi-cell RAN slicing problem from a comprehensive perspective incorporating these two aspects. A few previous studies, such as [5], can be found to consider slicing in multi-cell networks, despite considering only the traffic dimension and neglecting radio-electrical isolation analysis. Moreover, this paper further contributes to recent progress in research projects, which have to date mainly established the requirements and architectural context.

In greater detail, the analysis of RAN slicing is conducted in the next section in relation to the Radio Resource Management (RRM) functionalities that can support the split of radio resources among slices. Accordingly, four RAN slicing approaches are presented. The analysis is completed by comparing among the alternatives considering different perspectives, such as the granularity in the assignment of radio resources to the different tenants and the degrees of isolation and customization. Simulation-based evaluations further quantify the behavior of different slicing approaches in terms of isolation. Finally, conclusions and research directions are summarized.

RAN SLICING

From a radio resource perspective, let us consider a given amount of spectrum resources and a RAN composed by N cells deployed over a certain geographical area. Without loss of generality, let us assume that the spectrum resources can be flexibly arranged in a set of carriers (i.e., channels), with the same or different carrier bandwidth, and that the time/frequency dimensions of each carrier are organized in Resource Blocks (RBs), which constitute the basic physical radio resource unit used for the dynamic allocation of the radio capacity. For example, in the context of Long-Term Evolution (LTE) radio access, carrier bandwidths range from 1.4 to 20 MHz and a carrier of, e.g., 5 MHz is organized in 25 RBs of 180 kHz/0.5 ms each.

From a service perspective, the RAN provides Radio Access Bearers (RABs), which are the data delivery services offered for information exchange between the User Equipment (UE) and the mobile core network. In the context of LTE, a RAB is denoted as an Evolved-RAB (E-RAB) and is designed to transfer IP packets over the air interface; its expected behavior is parameterized with a set of Quality of Service (QoS) attributes (e.g., guaranteed bit rate, QoS class identifier, Allocation and Retention Priority) associated with the particular RAB and/or with the corresponding UE (e.g., maximum aggregated bit rate for all RABs of a UE).

On this basis, the deployment and exploitation of the RAN resources (cell sites and spectrum) to fulfill a given traffic demand (number and characteristics of the RABs) involve the following RRM functionalities:

- Spectrum planning: This function decides how the spectrum resources are arranged in carriers and how these carriers are assigned to the different cells.

- Inter-Cell Interference Coordination (ICIC): This function intends to mitigate the inter-cell interference that appears when neighbor cells use the same carrier. For this purpose, ICIC establishes a set of limitations on the usage of the different RBs, such as limiting the transmission power of a cell in certain RBs or forbidding the data transmission of a cell in certain RBs. Thus, from a general perspective, the ICIC strategy determines the set of RBs that each cell is allowed to use in each of its allocated carriers as well as the maximum transmission power for each RB.

- Packet Scheduling (PS): This function decides, for each carrier assigned to a cell, how the set of available RBs (i.e., those allowed by the ICIC strategy) are used to transfer the data traffic of the established RABs. The scheduling process operates at a resolution given by the so-called Transmission Time Interval (TTI), which currently is 1 ms in LTE and is expected to be even lower in 5G to better support ultra-low latency applications. The PS is also responsible for selecting the physical layer parameters used in the RB transmissions (e.g., modulation and coding scheme, antenna mapping in case of multi-antenna transmission). This selection determines the number of bits of each RAB that can be served in a TTI, thus controlling the bit rate (b/s) delivered to each RAB.

- Admission Control (AC): This function determines whether the establishment request of a new RAB is accepted or rejected in a given cell. The AC should account for the overall resource utilization in the cell, the QoS requirements of already active RABs and the requirements of the new RAB request.

It is worth noting that the achievable resource utilization efficiency of a specific RB (e.g., spectral efficiency in bit/s/Hz) depends on the physical layer techniques applied to the radio link (e.g., adaptive modulation and coding, hybrid automatic repeat request, multiple-input-multiple-output), the propagation conditions and the associated RRM decision-making processes (i.e., the aforementioned spectrum planning, ICIC, PS, AC), which affect the interference level observed at the air interface.

Let us assume that the RAN is to be shared among M tenants, denoted as T_1, ..., T_M, and that each tenant is provided with a RAN slice. According to the abovementioned RRM functionalities, different RAN slicing approaches are discussed in the following subsections.

A. RAN slicing at Spectrum Planning level

In this case, as illustrated in Fig. 1, RAN slicing is implemented by arranging the spectrum resources of the RAN infrastructure provider into a number of carriers and assigning different carriers to each tenant. The number and characteristics of the carriers assigned to a tenant should be sufficient to provide the required capacity and coverage over the entire scenario. Once particular carriers are assigned to a tenant, a tenant-specific spectrum planning function can decide how these carriers are used across the N cells, taking into account the specificities of the provided services and the spatial traffic distribution of the UEs. Moreover,

given that the separate carriers ensure both radio-electrical isolation and traffic isolation, not only the spectrum planning function but also the rest of mechanisms involved in the resource allocation to the RABs (i.e., ICIC, PS and AC) can be implemented according to tenant-specific policies. This situation is illustrated in Fig. 1 by depicting the RRM functions that can be particularized for tenant T_1 in red and those that could be specific for tenant T_M in green.

B. RAN slicing at ICIC level

In this case, as illustrated in Fig. 2, the spectrum planning decides the assignment of carriers to cells by taking into account the expected spatial distribution of the traffic aggregate and without assigning dedicated carriers per tenant. In this way, any specific carrier assigned to a cell will generally be shared among the tenants.

In its simpler implementation, this RAN slicing statically assigns a number of different RBs to each tenant in each carrier across all the cells in a certain area. This approach ensures that no interference exists between transmissions of different tenants in these cells, thus achieving both radio-electrical isolation and traffic isolation between slices. Once the RBs have been assigned to a tenant, a tenant-specific ICIC strategy can be used to establish the set of RBs of each carrier that the tenant can use in the different neighbor cells to mitigate the inter-cell interference within the RAN slice of the tenant. Therefore, thanks to the isolation between slices, different ICIC policies might be applied in each slice (e.g., one tenant implements a full reuse of its RBs in all the cells while another tenant implements a partial reuse). Similarly, tenant-specific PS and AC algorithms can be implemented. This situation is illustrated in Fig. 2 by depicting the functions that can be particularized for tenants T_1 and T_M in red and green, respectively, whereas the spectrum planning function, which is common for all the tenants, is depicted in blue.

It is worth noting that this approach follows a principle similar to "RAN slicing at Spectrum Planning" but with higher granularity because the slicing is performed over the RBs instead of over the carriers. It is also worth noting that other more sophisticated possibilities than the abovementioned static assignment could be devised by relying on the capability of ICIC algorithms to dynamically coordinate the use of the RBs across the mutually interfering cells and thus to dynamically modify the assignment of RBs to tenants in these cells in a coordinated manner.

C. RAN slicing at Packet Scheduling level

This approach is illustrated in Fig. 3. In this case, neither the spectrum planning nor the ICIC strategies assign dedicated radio resources (carriers and/or RBs) to the tenants across multiple cells. Instead, the RAN slicing is implemented at each cell by distributing the RBs among the tenants. Then, by limiting the number of RBs that can be used by each tenant, traffic isolation is ensured inside each cell. However, inter-cell

interference among tenants may arise because the ICIC function does not make distinctions among tenants when deciding the RBs that can be used in a cell.

Because of the traffic isolation provided inside a cell, this approach enables the application of tenant-specific PS algorithms to decide how the capacity provided by the RBs of the slice is distributed among its admitted RABs (e.g., different prioritization criteria to serve the RABs could be used for each tenant in high load conditions). In turn, spectrum planning and ICIC functions are common to all the tenants; thus, they are represented in blue in Fig. 3.

D. RAN slicing at Admission Control level

In this case, as illustrated in Fig. 4, RAN slicing is implemented by acting on the RAB admission process at each cell. A RAB requesting admission first goes through a tenant-specific AC that decides whether it can be admitted (e.g., a tenant may implement a different allocation and retention strategy, giving priority or pre-empting specific types of RABs when capacity limitations per slice are reached). Then, the RAN slicing function makes the final decision of acceptance or rejection considering the overall situation of all the tenants (e.g., requirements of all the RABs admitted and overall resource consumption) to ensure a proper split of capacity.

As in the RAN slicing at the PS level, radio-electrical isolation is not guaranteed because inter-cell interference between tenants may exist. Moreover, the PS function in a cell is common to all the tenants (i.e., it is depicted in blue in Fig. 4); thus, it distributes the available capacity among all admitted RABs without enforcing any tenant-specific treatment. For this reason, the AC executed by the RAN slicing is the key control mechanism adopted to avoid mutual effects between the traffic of different tenants and thus achieve a certain degree of traffic isolation.

ANALYSIS OF THE RAN SLICING APPROACHES

The analysis of the proposed RAN slicing approaches should take into account different perspectives, as summarized in Table I.

The first aspect to consider is the granularity achieved in the assignment of the radio resources to the different tenants. Granularity should be considered across the frequency dimension (i.e., the minimum amount of bandwidth that can be assigned to a tenant), the time dimension (i.e., the minimum amount of time that is needed before the assignment of resources to a tenant can be modified) and the space dimension (i.e., the minimum geographical region over which a resource assignment to a tenant applies). Granularity directly affects the flexibility to modify the resource assignments to adapt to different conditions. High flexibility is considered a positive feature in the sense that it provides the potential to assign more fine-grained resources to the different tenants and, therefore, achieves better utilization of these resources.

Focusing on the frequency dimension, the granularity of "RAN slicing at spectrum planning" is one carrier (e.g., 1.4 MHz in LTE). For both "RAN slicing at ICIC" and "RAN slicing at PS", the minimum frequency-domain unit that can be assigned to a tenant is one RB (e.g., 180 kHz in LTE). Regarding "RAN slicing at AC", no frequency granularity exists because the slicing is not applied over the physical radio resources but over the RAB establishment requests.

In the time dimension, "RAN slicing at spectrum planning" could change the carriers assigned to the tenants, which involves re-planning the network to modify the assignments of carriers to cells via dynamic spectrum allocation strategies [12]. Typically, such re-planning processes operate on a relatively long-term basis, at least on the order of several minutes. "RAN slicing at ICIC" involves modifications not in the carriers assigned to the cells but only in the distribution of the RBs of these carriers among the tenants. Therefore, in this case, the time granularity is constrained by the time scale of operation of the ICIC, which typically is on the order of hundreds of ms [13]. In turn, for "RAN slicing at PS", the time granularity of the RAN slicing is determined by the PS that operates at the TTI level, i.e., 1 ms in LTE (possibly even lower in 5G). Finally, for "RAB slicing at AC", the capability of dynamically modifying the split of capacity for each tenant is associated to the time scale of operation of the AC, which is executed each time a new RAB establishment request arrives to the system. Therefore, the time scale can be on the order of seconds.

The analysis of the space domain granularity is related to the geographical scope over which the RAN slicing to tenants can be modified. For "RAN slicing at PS" and "RAN slicing at AC", because slicing is performed at the cell level, different splits of resources can be performed at each cell, thus providing high flexibility. Instead, for "RAN slicing at spectrum planning" or "RAN slicing at ICIC", because both involve the consideration of interference levels among cells, the space granularity will expand over the whole service area (or perhaps over smaller portions of this service area if these portions do not mutually interfere).

The granularity in the time/frequency/space dimensions affects the capability of the different RAN slicing approaches to adapt ahead of the varying traffic conditions of the different tenants and to address situations in which some QoS requirements cannot be fulfilled due to resource limitations. Granularity is particularly useful in cases in which the resource limitations occur sporadically or in specific geographical areas affecting a reduced number of cells because the problem may be handled by temporarily increasing the amount of resources (e.g., RBs, carriers) assigned to the affected tenant(s), at the expense of reducing the amount of resources assigned to other tenants that are not consuming all their resources. Similarly, it may be possible to increase the resources assigned to a tenant in some affected cells at the expense of reducing the assignment in other cells where the tenant has less traffic. Instead, if QoS requirements of tenants are not met for long periods and/or affect a large number of cells, a network dimensioning problem is implied, which must be addressed by the network infrastructure provider by, e.g., deploying more cells or assigning more spectrum to

the deployed cells.

Another aspect to consider is the capability of each technique to customize the RRM strategies on a per-tenant basis. The highest degree of customization is provided by "RAN slicing at spectrum planning" that allows implementing tenant-specific policies for spectrum planning, ICIC, PS and AC. In turn, "RAN slicing at ICIC" allows for the customization of ICIC, PS and AC, whereas "RAN slicing at PS" allows for the customization of PS and AC on a per-tenant basis. The lowest customization is obtained with "RAN slicing at AC", in which only AC can be tenant-specific. A high level of customization can be viewed as a positive feature in the sense that it provides the potential to apply RRM strategies suitably adapted to the specific and diverse requirements of each tenant.

Table I also presents a qualitative comparison in terms of the level of isolation achieved with each approach. The highest radio-electrical isolation is obtained with "RAN slicing at spectrum planning" because in this case each tenant will use different carriers over the entire scenario. Therefore, interference among tenants can only appear in the form of adjacent channel interference, which typically leads to Carrier to Interference Ratio (CIR) values well above 30 dB. For "RAN slicing at ICIC", high isolation is also achieved because transmissions of different tenants in different cells will use different RBs. Therefore, no inter-cell interference among tenants will exist. Only adjacent channel interference or interference between adjacent subcarriers inside the same carrier due to imperfect synchronization and the Doppler effect will be present, which again leads to high CIR values typically above 25 dB.

For "RAN slicing at PS" and "RAN slicing at AC", inter-cell interference among tenants can potentially appear, depending on how the ICIC and spectrum planning strategies have distributed the carriers and RBs among the cells. For example, if the ICIC decides to reuse an RB in two neighboring cells, this RB can be used by a different tenant in each cell, leading to inter-cell interference between both tenants. Instead, if the ICIC decides that an RB is not reused in two neighboring cells, no inter-cell interference between tenants will exist.

Traffic isolation is ensured with RAN slicing at the spectrum planning, ICIC or PS level because the PS of a tenant can only use, at maximum, the carriers and/or RBs that the slicing process has assigned to this tenant. Therefore, an overload of this tenant will only affect the performance experienced by the RABs of this tenant but not the RABs of the others. Instead, when RAN slicing is done at AC, the PS does not distinguish among the RABs of the different tenants when distributing the available capacity. Therefore, the isolation relies on the capability of AC and RAN slicing to restrict the admitted RABs of each tenant to prevent a high-load situation for one tenant from affecting other tenants. However, due to the high variability of data traffic and the fact that AC typically operates based on statistical estimations of the resource consumption of the

RABs, overload situations cannot be completely avoided. Therefore, this approach achieves the lowest traffic isolation among those considered and depends on how restrictive the AC is.

QUANTITATIVE ANALYSIS OF THE ISOLATION

In addition to the qualitative comparison presented in Table I, this section presents a simulation-based analysis to quantify the differences in terms of isolation that can be achieved between different techniques.

Simulations consider a scenario in which the RAN is sliced to provide access services to two tenants (denoted as A and B) in two cells. The distance between the cell sites is 1.4 km. The propagation losses in dB are 128.1 + 37.6 log d + S, where d(km) is the distance and S(dB) is a Gaussian random variable with 0 mean and standard deviation σ(dB) representing the shadowing. The downlink direction is considered, and the noise power per RB is -115.44 dBm.

The UEs of the two tenants are uniformly distributed in the scenario. They generate data sessions whose duration is exponentially distributed with an average of 30 s. Session generation follows for a Poisson process in which the session rate per cell for tenant A is 0.15 sessions/s, whereas the session rate of tenant B is varied in the different simulations to represent different load conditions. Each session of a UE requires the establishment of a RAB with a required bit rate $R_{b,req}$=1024 kb/s. Then, the PS will assign to each UE the required number of RBs to fulfill the bit rate $R_{b,req}$ depending on the Signal to Interference and Noise Ratio (SINR) experienced by the UE. The mapping between the bit rate achievable in an RB and the SINR is computed based on the model presented in section A.1 of [15].

Each cell has two carriers allocated, f_1 and f_2, each one with 25 RBs. The total transmitted power available per carrier is initially set to P=37 dBm. Two possible options are considered for ICIC, namely a Full Frequency Reuse (Full FR) and a Fractional Frequency Reuse (FFR) ICIC strategy [14]. In Full FR, all the RBs of a carrier are used by the two cells and the total transmitted power P is equally split among the 25 RBs (i.e., the transmitted power per RB is 23 dBm). In FFR, the 25 RBs of a carrier are split between 9 inner RBs and 16 outer RBs. The inner RBs are used at both cells and assigned to inner UEs (i.e., UEs at distance below 500 m). The outer RBs are further split into two equal sets of 8 RBs, each one used at a different cell. Outer RBs are assigned to outer UEs (i.e., UEs at distance above 500 m). For FFR, the total transmitted power in one carrier is equally split among the 17 RBs (9 inner + 8 outer) assigned to the carrier such that the transmitted power per RB is 24.7 dBm.

Radio-electrical isolation

The analysis in this section focuses on “RAN slicing at spectrum planning” and “RAN slicing at PS” approaches as representative of what “high” and “medium” radio-electrical isolation levels represent in Table I.

For “RAN slicing at spectrum planning”, carrier f_1 is assigned to Tenant A and carrier f_2 is assigned to Tenant B. However, “RAN slicing at PS” dynamically distributes the RBs of the two carriers in a cell among the two tenants. For a fair comparison with “RAN slicing at spectrum planning”, the criterion for performing this distribution is to assign 50% of the total RBs in a cell to each tenant (i.e., N_{RB}=25 RBs for each tenant in the Full FR case and N_{RB}=17 RBs in the FFR case). However, if the PS of a tenant is able to fulfill the bit rate requirements of its currently active RABs with fewer RBs, the remaining RBs can be assigned to the other tenant. In both slicing approaches, the AC function is executed separately for each tenant. The AC criterion assumes that a new RAB is admitted if the average number of RBs used by the tenant after acceptance of the RAB is estimated to be lower than a threshold set to 95% of N_{RB}. This estimation is based on measurements of the average RB usage performed during the simulation.

To quantify the radio-electrical isolation, Fig. 5 plots the average SINR per RB experienced by the users of Tenant A as a function of the offered load per cell of Tenant B for the two slicing approaches considered. Fig. 5a presents the Full FR case for two different values of the shadowing standard deviation σ. It can be observed that for “RAN slicing at spectrum planning” the SINR is not affected by Tenant B’s load conditions because the slices operate over separated carriers. Instead, with “RAN slicing at PS”, the two carriers are shared; thus, the SINR of the Tenant A’s UEs is reduced when increasing the load of Tenant B because this situation leads to a higher utilization of the RBs in the neighboring cell and thus to higher inter-cell interference. Differences of up to 3 dB in SINR are observed by UEs of Tenant A between the lowest and the highest load levels of Tenant B, reflecting the lower radio-electrical isolation provided by “RAN slicing at PS”. The same behavior is observed for both values of σ. Indeed, this lower radio-electrical isolation (i.e., lower SINR) implies that, for high load conditions of Tenant B, more RBs are required by each UE of Tenant A to achieve the requested bit rate $R_{b,req}$. Although not shown here for brevity, simulations revealed that the 3 dB reduction in SINR of Fig. 5a results in an increase of 25% in the average number of RBs required by each UE. In contrast, it is worth noting that, for a low load of Tenant B, the SINR experienced by the UEs of Tenant A is better for “RAN slicing at PS level”, leading to better utilization of the radio resources because fewer RBs will be needed to satisfy $R_{b,req}$.

Fig. 5b presents the FFR case with σ=6 dB. In addition to the average SINR considering all the users of Tenant A, the figure also shows the average SINR of the inner and outer users. In absolute terms, higher average SINR values than in Full FR are observed because of lower interference. For the inner users, an effect similar to that noted for Full FR is observed, i.e., reduction of SINR by approximately 3 dB when increasing the load of Tenant B with “RAN slicing at PS”. In contrast, it is observed that the SINR of outer users is not influenced by the load of Tenant B in this case, regardless of the considered slicing approach, because, even with “RAN slicing at PS”, the FFR ICIC strategy ensures that no inter-cell interference will

be experienced by outer users, thus providing radio-electrical isolation. Due to this different behavior of the inner and outer users, the average SINR considering all users exhibits a lower reduction (i.e., 1.3 dB) when increasing the load of Tenant B than in the Full FR case of Fig. 5a.

Traffic isolation

Traffic isolation is analyzed by comparing the "RAN slicing at PS" and "RAN slicing at AC" approaches as representative of what "high" and "medium" isolation levels indicate in Table I. The "RAN slicing at AC" strategy assumes that the AC accounts for both the tenant-based admission condition discussed in the previous subsection plus a second condition to ensure that the aggregate average number of RBs used by the two tenants in a cell does not exceed the 95% of total RBs available in the cell.

To quantify the effects of the traffic isolation, Fig. 6 depicts the probability that the bit rate R_b experienced by a UE of Tenant A is below the requirement $R_{b,req}$ as a function of the offered load of Tenant B. This degradation in the experienced bit rate occurs under overload conditions, i.e., whenever the PS does not have sufficient RBs to serve the admitted RABs, typically because the number of RBs required by each UE to achieve $R_{b,req}$ is not fixed but varies with the propagation and interference conditions of each UE; thus, the figure cannot be perfectly estimated at the AC stage.

Focusing on "RAN slicing at PS", Fig. 6 shows that, when the load of Tenant B is very small, the probability that the UEs of Tenant A do not achieve $R_{b,req}$ is close to 0. The reason is that, in such a case, Tenant B only requires a very small number of RBs. Therefore, the RAN slicing can assign more RBs to Tenant A, and accordingly, the PS of this tenant will have sufficient RBs to serve its UEs in most cases. As the load of Tenant B increases, more RBs will be used by this tenant; therefore, Tenant A can use fewer RBs, and the probability of not achieving $R_{b,req}$ starts to increase. However, for very high loads of Tenant B, this probability tends to saturate to a constant value (i.e., approximately 0.7% for the case P=37 dBm and 3.3% for the case P=30 dBm in Fig. 6). This saturation occurs because "RAN slicing at PS" always ensures that Tenant A can use at least N_{RB}=25 RBs to serve its users, regardless of the load of Tenant B. This finding reflects the traffic isolation achieved by this slicing technique that prevents the performance of Tenant A from degrading beyond the saturation limit even when the load of Tenant B increases. Instead, with "RAN slicing at AC", Fig. 6 shows that the performance experienced by Tenant A users suffers greater degradation when increasing the load of Tenant B. The reason is that, in this case, the only limits on the load of each tenant are set by the blockings performed by the AC, whereas the PS is not tenant-specific and thus cannot guarantee a minimum number of RBs per tenant. Therefore, this situation results in a worse traffic isolation.

CONCLUSIONS AND RESEARCH DIRECTIONS

Network slicing, in which a common network infrastructure is exposed in the form of different customized logical networks, each one tailored to support specific applications or to be operated by a different communication provider, has become a fundamental concept in addressing the heterogeneity of performance requirements and network functionalities associated with the multiple use cases of the 5G ecosystem. This paper has analyzed the application of the network slicing concept to a multi-cell RAN that is shared among multiple tenants. Due to the characteristics of the radio channel and the potential effect that any transmitter may have on any receiver, the slicing of a multi-cell RAN is a challenging process because it requires considering isolation among tenants from two different perspectives, namely traffic isolation and the radio-electrical isolation. Accordingly, the paper has addressed the RAN slicing problem from a comprehensive perspective incorporating these two concepts and has presented four possible RAN slicing approaches that differ in the RRM functions used as a support for splitting the radio resources among slices.

The presented alternatives have been compared along different dimensions, both qualitatively and quantitatively. The highest level of radio-electrical isolation is obtained with the "RAN slicing at spectrum planning" and "RAN slicing at ICIC" approaches, which prevent co-channel inter-cell interference among tenants, unlike "RAN slicing at PS" or "RAN slicing at AC", in which potential inter-cell interference among tenants may appear. However, the latter two techniques offer higher granularity and more flexibility in the assignment of radio resources to tenants. In turn, the lowest level of traffic isolation is provided by "RAN slicing at AC", whereas the other approaches ensure higher isolation because they limit the number of RBs that can be used by each tenant. The different approaches also offer different degrees of customization among tenants because they establish the RRM functions that have to be common to all tenants and those that can be implemented following tenant-specific policies. In this respect, the highest customization is provided by "RAN slicing at spectrum planning", which allows the spectrum planning, ICIC, PS and AC functions to be specified on a per-tenant basis. In contrast, the lowest customization is obtained with "RAN slicing at AC", in which only the AC can be made tenant-specific. Overall, although the RAN slicing approaches providing high isolation might be deemed superior, qualitative aspects such as flexibility and customization as well as their corresponding implementation complexity may advocate for the realization of other solutions.

Following the characterization of the multi-cell RAN slicing problem, different research challenges are identified for contributing to the consolidation of this concept in future 5G networks. In particular, the RAN slicing should be materialized in specific and efficient algorithmic solutions capable of fulfilling the capacity and QoS requirements of each tenant, which are typically associated with established Service Level Agreements between the infrastructure provider and the tenant. To this end, for the RAN slicing approaches executed at the cell level, appropriate coordination mechanisms among cells should be devised to ensure that

the capacity is consistently distributed across the whole RAN ahead of non-homogeneous spatial traffic distributions of the different tenants. Finally, the possibilities implied by Software Defined Network (SDN) and Network Function Virtualization (NFV) technologies for implementing each RAN slicing approach should be analyzed. Specifically, advancing towards programmable control of RRM functions within the RAN together with the possibility to run per-tenant RRM functions as Virtual Network Functions (VNFs) appears to be a flexible framework that may facilitate the materialization of diverse business and network operation models.

ACKNOWLEDGEMENTS

This work has been supported by the EU funded H2020 5G-PPP project SESAME under the grant agreement 671596 and by the Spanish Research Council and FEDER funds under RAMSES grant (ref. TEC2013-41698-R).

REFERENCES

[1] NGMN Alliance, "5G White Paper", February, 2015.

[2] 3GPP TS 23.251 v13.1.0, "Network Sharing; Architecture and functional description (Release 13)", March, 2015.

[3] R. Kokku, R. Mahindra, H. Zhang, S. Rangarajan, "NVS: A substrate for Virtualizing Wireless Resources in Cellular Networks", IEEE/ACM Transactions on Networking, Vol. 20, No. 5, October, 2012.

[4] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, "Radio Access Network Virtualization for Future Mobile Carrier Networks", IEEE Communications Magazine, July, 2013.

[5] R. Mahindra, M. Khojastepour, H. Zhang, S. Rangarajan, "Radio Access Networks Sharing in Cellular Networks", 21st IEEE International Conference on Network Protocols (ICNP), 2013

[6] 3GPP TR22.891 v2.0.0 "Feasibility Study on New Services and Markets Technology Enablers; Stage 1 (Release 14)", February, 2016.

[7] P.Rost (editor), "Functional Network Architecture and Security Requirements", Deliverable D3.1 of the 5G NORMA project, December, 2015.

[8] METIS II White Paper "Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design", March, 2016

[9] J. O. Fajardo, et al, "Introducing mobile edge computing capabilities through distributed 5G Cloud Enabled Small Cells", Accepted at ACM/Springer Mobile Networks and Applications (MONET), 2016.

[10] V. Del Piccolo, A. Amamou, K. Haddadou, G. Pujolle, "A Survey of network isolation solutions for multi-tenant data centers", IEEE Communications Surveys and Tutorials, IEEE Early Access Articles, 2016.

[11] C. Liang, F. R. Yu, “Wireless Network Virtualization: A survey, some research issues and challenges”, IEEE Communications Surveys and Tutorials, Vol. 17, No. 1, 1st Quarter, 2015.

[12] K. Spaey et al. “SON functions for Multi-layer LTE and Multi-RAT networks (Final Results)”, Deliverable D4.2 of the SEMAFOUR project, August, 2014, available at http://www.fp7-semafour.eu/en/public-deliverables/

[13] J. Ramiro, K. Hamied, Self-Organizing Networks. Self-planning, self-optimization and self-healing for GSM, UMTS and LTE, John Wiley & Sons, 2012.

[14] T. D. Novlan, R. K. Ganti, A. Ghosh, J. G. Andrews, “Analytical Evaluation of Fractional Frequency Reuse for OFDMA Cellular Networks”, IEEE Transactions on Wireless Communications, Vol. 10, No. 12, December, 2011, pp. 4294-4304.

[15] 3GPP TR 36.942 v12.0.0, “Radio Frequency (RF) system scenarios”, September, 2014.

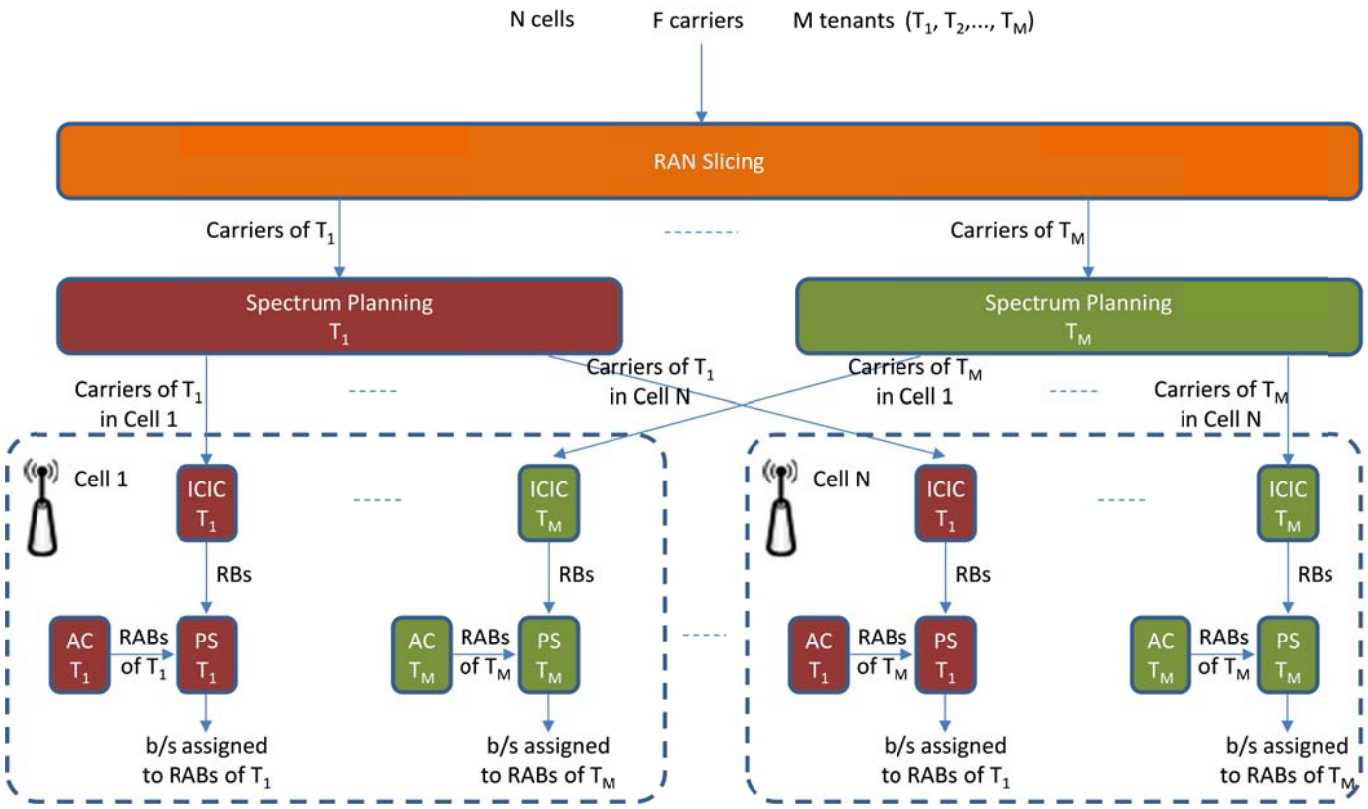


Fig. 1. RAN slicing at Spectrum Planning level

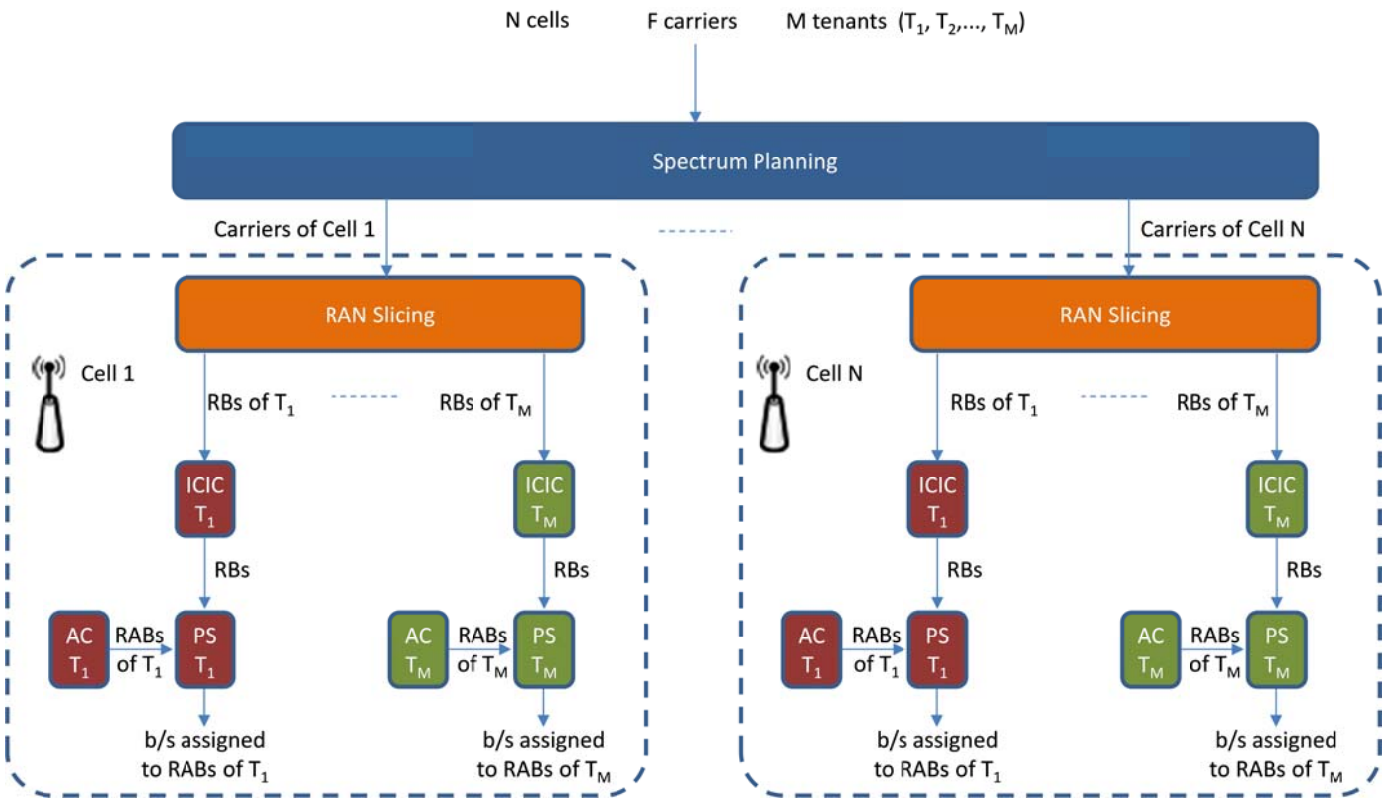


Fig. 2. RAN slicing at ICIC level

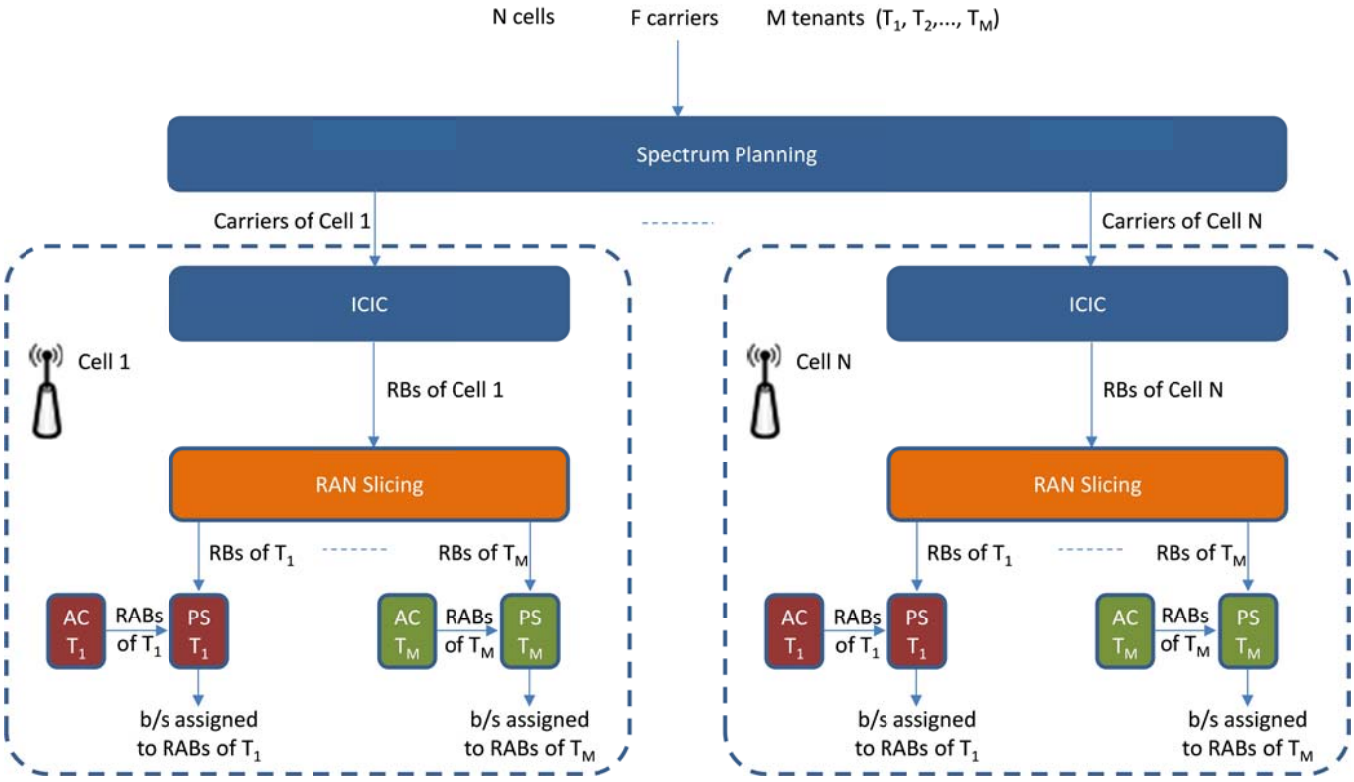


Fig. 3. RAN slicing at Packet Scheduling level

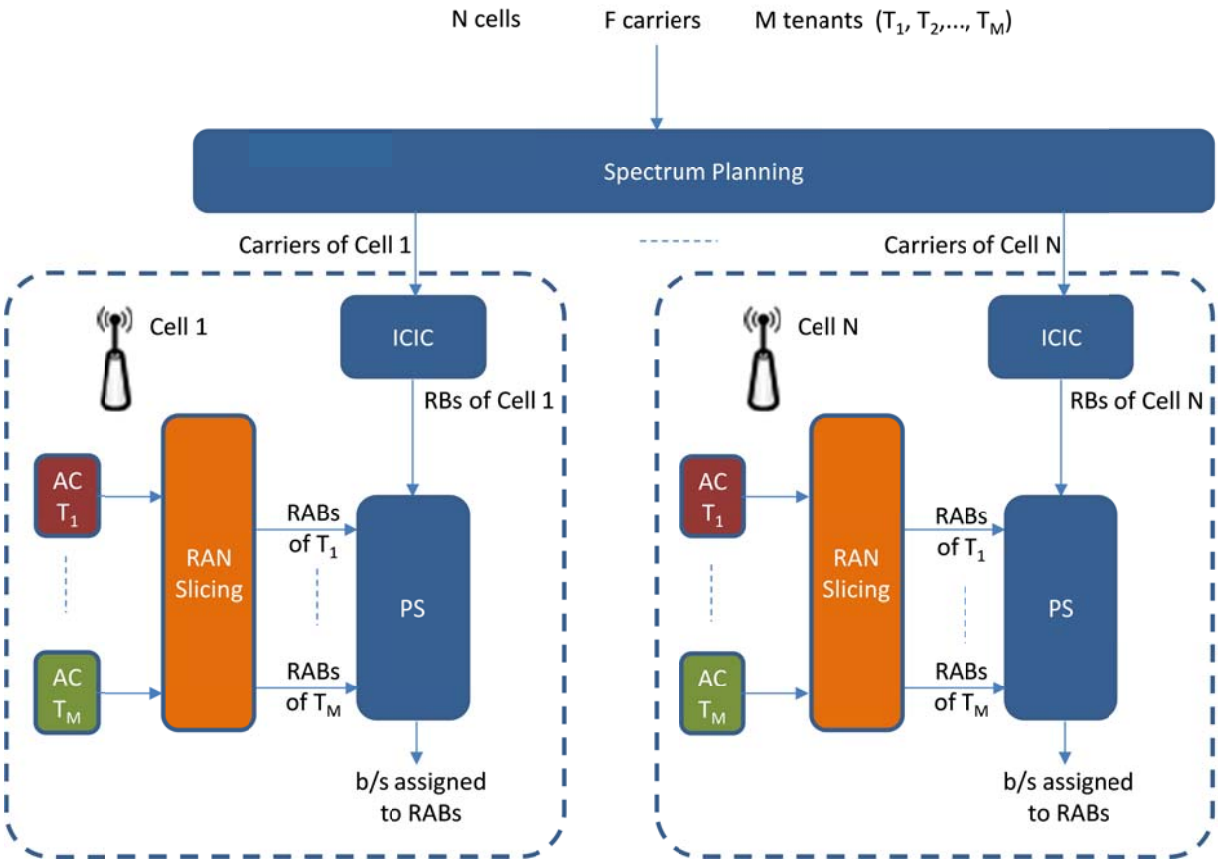
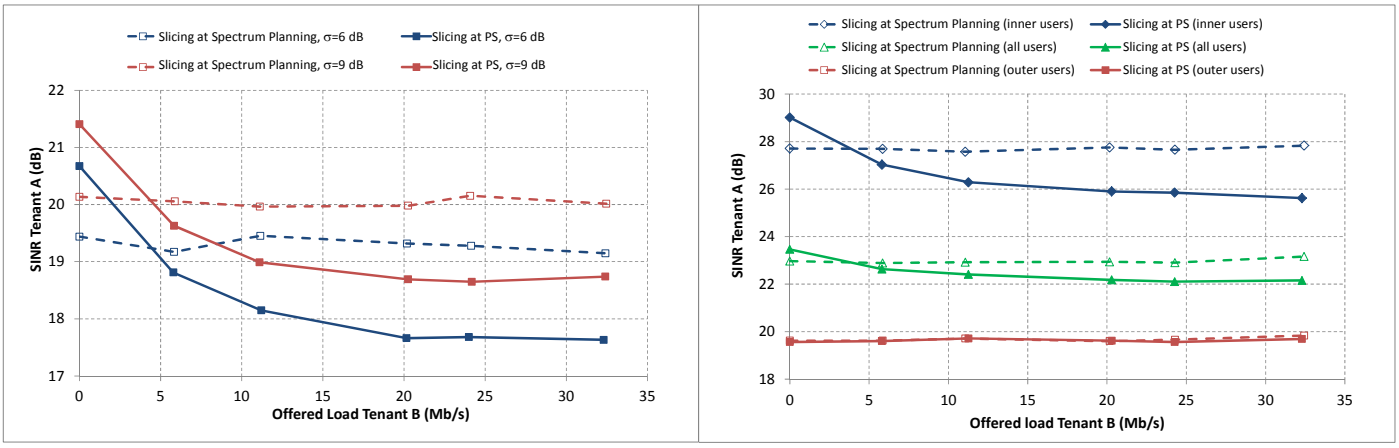


Fig. 4. RAN slicing at Admission Control level



(a) (b)

Fig. 5. Average SINR experienced by the users of Tenant A as a function of the offered load per cell of Tenant B for "RAN slicing at spectrum planning" and "RAN slicing at PS" approaches. (a) Full FR case and (b) FFR case.

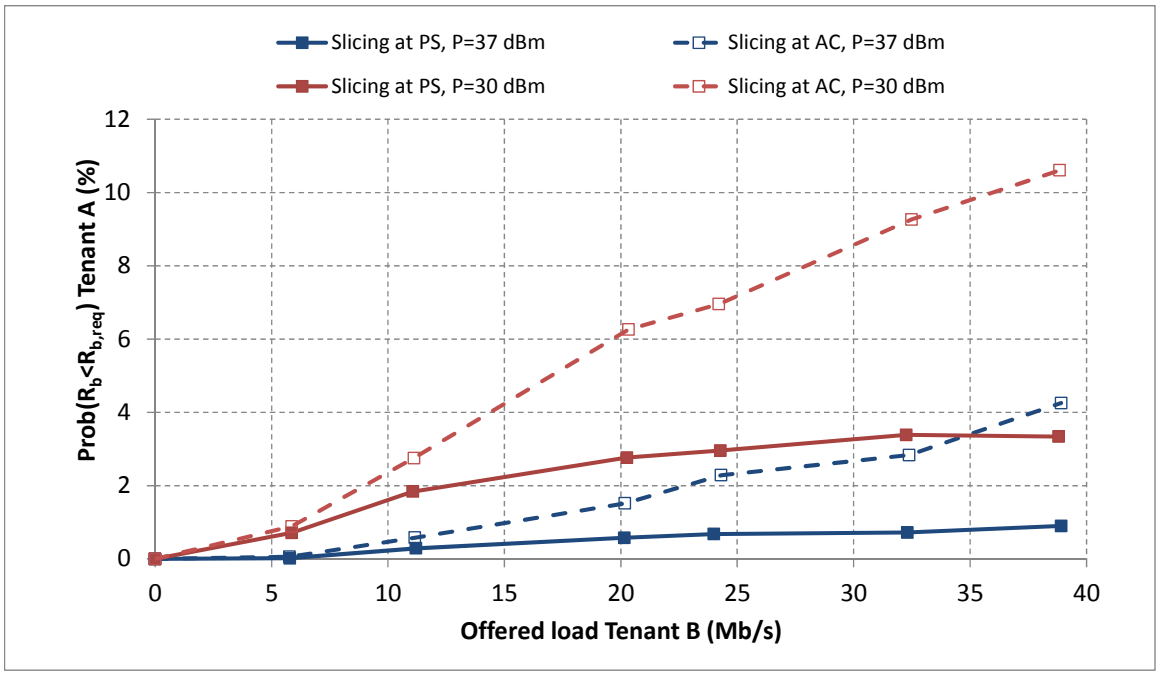


Fig. 6. Probability that the bit rate experienced by a user of tenant A is below the required bit rate as a function of the offered load per cell of Tenant B for the "RAN slicing at PS level" and "RAN slicing at AC level" approaches. Simulations consider the Full FR case, σ=6 dB and two different transmitted power levels per carrier P=30 dBm and P=37 dBm.

Table I. Comparison between RAN slicing strategies

	RAN slicing at Spectrum Planning level	**RAN slicing at ICIC level**	**RAN slicing at Packet Scheduling level**	**RAN slicing at Admission Control level**
Granularity in the frequency domain	1 carrier (minimum 1.4 MHz in LTE)	1 RB (180 kHz in LTE)	1 RB (180 kHz in LTE)	Not applicable
Granularity in the time domain	Relatively long-term (several minutes)	Every ICIC period (typically hundreds of ms in LTE)	Every TTI (1ms in LTE)	Associated to RAB establishment request rate (in the order of seconds)
Granularity in the spatial domain	Whole scenario (might be less if sets of non-interfering cells are identified)	Whole scenario (might be less if sets of non-interfering cells are identified)	One cell	One cell
Degree of customization	Spectrum planning, ICIC, PS and AC can be tenant-specific	ICIC, PS and AC can be tenant-specific	PS and AC can be tenant-specific	AC can be tenant-specific
Radio-electrical isolation	High	High	Medium	Medium
Traffic isolation	High	High	High	Medium