5G RAN slicing for verticals: Enablers and challenges

Salah Eddine Elayoubi, Sana Benjemaa, Zwi Altman, Ana Maria Galindo Serrano

5G RAN slicing for verticals: Enablers and challenges

Salah Eddine Elayoubi, Sana Ben Jemaa, Zwi Altman, Ana Galindo-Serrano

Abstract— This article investigates the slicing concept in the 5G Radio Access Network (RAN) with the related challenges and research problems. The objective is to identify the plausible options for implementing the slicing concept at the RAN level by the Mobile Network Operator (MNO) to respond to the needs of verticals. We start by identifying the different slice granularity options, i.e., how to define slices by combining customer and service needs. We then present how the 5G New Radio (NR) features can be used for facilitating slice implementation and provide typical configurations for different slice types from technology and RAN architecture perspectives. The main challenges for RAN slicing are then discussed, with a special attention to the resource allocation problem between slices sharing the same spectrum band. We also investigate the multi-tenant slicing implementation in terms of the openness of the network to third parties which is regarded as a key issue that may encourage vertical players to use operators' networks rather than deploying their own infrastructure.

Index Terms—Mobile networks, 5G, Slicing, Network design

1 Introduction

While third and fourth generation mobile networks revolutionized social behaviors by enabling the generalization of social networking on mobile devices, fifth Generation (5G) networks are expected to revolutionize our living environments, our cities and our industry by connecting everything. 5G is thus expected to mark a disruptive change and not to be a business-as-usual evolution of 4G networks limited to higher user throughputs but has to cover the needs of smart cities and vertical industries. In addition to the so-called enhanced Mobile Broadband (eMBB) service, the 5G design has to meet the requirements of two "new" mobile services: massive Machine Type Communications (mMTC), characterized by a very large density of connected objects and Ultra-Reliable Low Latency Communications (URLLC), characterized by stringent requirements in terms of low latency and high reliability. This is illustrated in Figure 1 along with some associated vertical use cases.

Considering each service type separately and building a network accordingly would likely end up with very different Radio Access Network (RAN) designs and architectures. However, only a common RAN that accommodates all three service types is an economically and environmentally sustainable solution. In addition, 5G is viewed not only as a new radio and core, but also as an orchestration platform where verticals build specialized services for their customers **[1]**. This creates a large number of services, that belong to the above-defined three service families (eMBB, URLLC and mMTC) **[2]**, but with a plethora of requirements. The slicing concept has then emerged as an efficient way for serving all these services on a common infrastructure [3]. A slice may describe an end-to-end relationship, i.e., its functionalities may also cover the 5G RAN. A network slice can be considered as an independent network, with the corresponding advantages e.g., for security and guaranteed Service Level Agreement (SLA). However, in contrast to deploying independent network infrastructures as it was the case in former mobile radio generations, the slices may be implemented, completely or partially, on a common infrastructure layer, including assets such as spectrum.

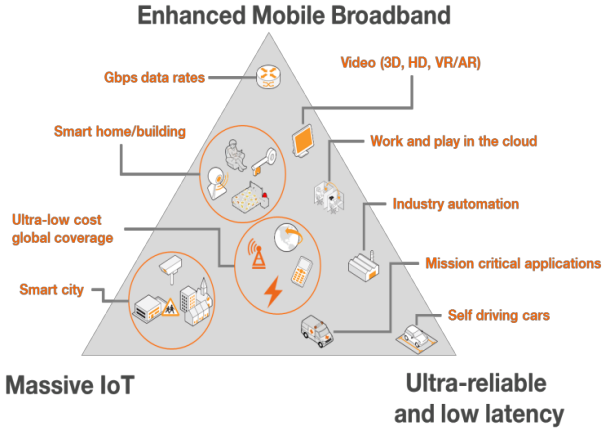


Fig 1. 5G as enabler for new services

While the utility of slicing for serving vertical use cases is commonly understood, there is still no consensus on the slice granularity nor the slice implementation on the RAN. In **[8]**, the authors identify the different requirements for slicing in the RAN, including efficiency, protection, differentiation and slice-awareness. **[3]** discusses the management and orchestration for end-to-end slices, including infrastructure layer, network function layer and service layer, but does not enter into the RAN implementation details. **[5]** Focuses on the architectural dimension for slicing, including network function chaining, leveraging on the concept of Software Defined Mobile Network Control (SDN-C).

This paper discusses how the RAN can be sliced for

- *Salah Eddine Elayoubi is with CentraleSupélec, L2S laboratory, 3 rue Joliot Curie, 91190 Gif-Sur-Yvette, France. E-mail: salaheddine.elayoubi@ centralesupelec.fr.*
- *Sana Ben Jemaa, Zwi Altman and Ana Galindo-Serrano are with Orange Labs, 44 avenue de la République, 92230 Châtillon, France. E-mails: firstname.lastname@orange.com.*

satisfying heterogeneous service requirements while sharing the same radio and processing resources. For doing so, we show how 5G new radio features, like tiling and puncturing, can be exploited and how different functions can be placed and chained using practical examples. In particular, we provide suggestions on how the network functions on lower layers can be configured for the different slices and how the architecture has to be adapted so that the service performance targets are met. We also give a special emphasis on the resource allocation techniques that are suitable for different use cases, including scheduling and channel access, and on how the corresponding slices interact among them at the RAN level. We provide discussions on the relevant challenges and corresponding open research problems.

2 SLICING GRANULARITY AT THE RAN

Before entering into the different options for defining a slice, we note that slicing has also been imagined as a mean for simplifying and optimizing network and infrastructure sharing between operators. In this context, one slice could be associated with an operator's virtual network deployed on another operator's infrastructure. We note that slicing for dynamic sharing of infrastructure and/or spectrum between operators is out of the scope of this paper.

Option 1: one slice per family of services

The simplest way for defining slicing is to consider a slice per family of services, for example one for smartphones (eMBB service), one for autonomous driving (URLLC service) and one for the Internet of Things (IoT, mMTC service). However, this approach does not consider the heterogeneity of requirements within each family of services. For example, the URLLC service includes a large set of use cases with very different requirements, ranging from wide area haptic services with very low end-to-end latency and low mobility requirements, to local networks of moving robots with stringent reliability constraints.

Option 2: one slice per set of technical requirements

A more general approach is to define a small number of technical slices starting from use-case requirements (bandwidth, latency, security, volume of messages, scalability, mobility, etc.) and grouping services that belong to the same family (in terms of requirements). While this approach solves the above mentioned issue (heterogeneity of requirements within the same 5G service family), other issues arise. For instance, slices offering the same type of services to different players (e.g., Renault or PSA), with different SLAs still need to be differentiated and isolated.

Option 3: one slice per vertical customer

We now consider an alternative approach with a slice per vertical customer. This option does not, however, correspond to a clear definition of slices. For instance, if we consider a business customer from the automotive domain, we can identify several services with heterogeneous requirements:

- Entertainment (high throughput, close to eMBB),
- Mission critical/autonomous driving (low latency high availability close to URLLC),
- Data retrieving for maintenance or tracking (low throughput, close to mMTC).

These different services cannot be managed by a single slice properly given the heterogeneity of their requirements.

Option 4: one slice defined per business customer and technical requirements

The analysis of the previous three options calls for a large number of slices, defined on business and SLA bases. For example, this option defines a slice for a specific automotive customer for entertainment on board, another one for the same customer for autonomous driving, while another slice is dedicated to autonomous driving for another automotive customer, and so on. This option is coherent with the definition of "a network slice instance" given in the 3GPP standard [4]. In practice, the operator may propose to the customer (Vertical) a network slice (e.g., URLLC for automotive industry) with configurable characteristics. The deployed slice instance (URLLC for automotive industry of customer x) would correspond to the customization of this Network slice to respond to the specific agreed SLA. In 3GPP [4], a UE can select a specific slice knowing its Single Network Slice Selection Assistance Information (S-NSSI). This slice identifier consists of two components:

- A Slice/Service type (SST), which refers to the expected slice behaviour in terms of features and services. Three standardized SST values are defined in [4], namely eMBB, URLLC and mMTC in order to provide a way for establishing global interoperability for slicing (e.g. in case of roaming), but additional SSTs can be defined.
- A Slice Differentiator (SD), which is optional information that complements the SST to differentiate amongst multiple slices of the same Slice/Service type, in order to differentiate two slices corresponding to different customers, with different SLA requirements for example.

3 RAN SLICING IMPLEMENTATION

Starting from the granularity of slicing of option 4 above, we here identify how these slices are mapped and implemented on the RAN level in a cost-effective way in terms of radio resource consumption. We start from the lower layers up to network function selection, configuration and chaining for each slice.

3.1 Slice isolation and spectrum sharing

One of the main challenges in implementing slicing in the RAN consists in designing and managing several slices on the same shared infrastructure in an efficient manner, while guaranteeing the agreed SLA for each of them. This brings us to the "slice isolation" concept that forbids any mutual impact between coexisting slices. The isolation concept is not clear today when it comes to resource allocation (while it is easy to understand in other domains such as security), but it is of common understanding to consider that two slices are isolated as long as the actions performed on one slice do not result in an SLA violation on the other slice.

Depending on the way the spectrum is shared and the slices are multiplexed in the RAN, the degree of isolation of slices from a performance perspective varies [5]. At the extreme left of Figure 2, a slice corresponds to a stand-alone network, with its specific spectrum and infrastructure. On the opposite, at the extreme right side, slices are limited to core network and resource allocation in the RAN is slice-independent, possibly implementing 4G-like Quality of Service (QoS) differentiation mechanisms. Intermediate solutions are also possible, with slices going lower in the protocol stack when moving left. It is worth noting that customization and isolation increase when moving left, at the expense of higher needs in terms of infrastructure deployment.

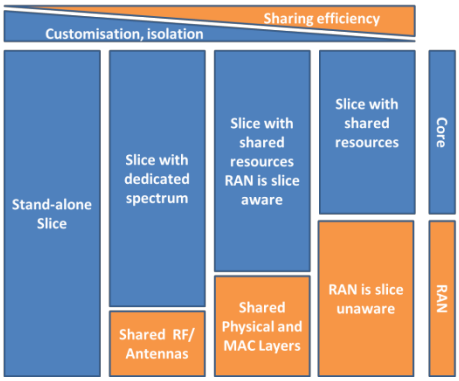


Fig. 2. RAN slice multiplexing options.

While neither of the slicing options in Figure 2 are to be discarded, the extreme cases correspond to very specific situations (a stand-alone network vs. a 4G-like RAN).
We propose a general slice implementation as an intermediate degree of isolation. In this proposed scheme, the physical (PHY) and Medium Access (MAC) layers common to all slices (except those on dedicated or unlicensed spectrum), and the higher layers, are slice specific. For these higher layers, a specific network function selection, configuration and chaining is to be performed for each slice. This chaining aims at achieving specific requirements (e.g., suppressing some processing functions for reducing latency). For example, handover management functions can be suppressed for slices serving fixed nodes, and packet fragmentation can be removed for slices transporting short but critical packets.

3.2 Tiling and scheduling

A first radio feature that is considered as an essential enabler for slicing at the RAN level is the tiling scheme [5]. This latter is a practical implementation of the so-called flexible numerology concept of 5G. Recall that resource allocation in 4G is based on a time-frequency grid, with a basic allocation of one Resource Block (RB) composed of 7 subcarriers of 15 KHz each allocated during a slot of 0.5 ms. 5G offers the opportunity of serving different services using different subcarrier spacing and/or Transmission Time Interval (TTI) lengths. The principle is that time-frequency resources with the same numerology are grouped together within a tile (or RB Group, RBG). This reduces the processing burden associated to scheduling as it restricts the positions where users of different services may be allocated and minimizes border issues (interference between different numerologies). Figure 3 illustrates this concept with the following design principles [6]:

- For eMBB: Start transport (TCP) sessions with short TTI size to quickly overcome the slow-start phase (third tile of Figure 3), then use a medium TTI (e.g., 1 ms, first tile of Figure 3) to minimize control overhead.
- For mMTC: use a lower bandwidth with a longer TTI size to save device energy and increase coverage (similar to the Narrow Band IoT concept of 4G). This is illustrated by tile 2 in Figure 3.
- For URLLC: use short TTIs (e.g., 0.25 ms) to meet latency requirements. Larger subcarrier spacing can also be useful for some URLLC use cases (e.g., in vehicular case against Doppler effect), see tile 3 of Figure 3.

It is worth noting that the tiling concept can be extended to mixed waveforms (e.g., CP-OFDM for eMBB and a filtered version for mMTC).

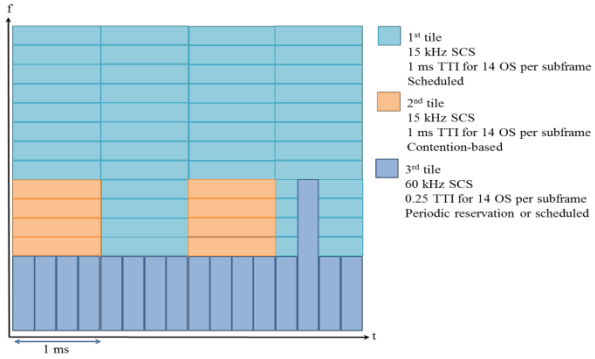


Fig. 3. Example of tiles. A first 4G-like is used for general eMBB services in scheduled mode. The second tile is allocated to mMTC services in a contention-based mode. The third tile with a smaller TTI and a larger SCS is suitable for URLLC transmissions but also for eMBB flows in their slow start phase.

We now explore the usage of the tiling concept for optimal resource allocation for slices. The scheduler has the role of allocating resources to comply to SLAs with the corresponding QoS requirements for the different slices, with possibly highly heterogeneous requirements. The scheduler complexity can be simplified by (i) dynamically determining the tile composition and (ii) by mapping the slices to these tiles. Using such a mapping, a simpler management of numerology and other PHY/MAC parameters such as TTI length and waveform parameters can be achieved.

We note that multiplexing in time is easy as 3GPP imposes symbol alignment between tiles, ensuring orthogonality. As of frequency multiplexing, in order to guarantee orthogonality between adjacent tiles, 3GPP advocates the insertion of guard bands between tiles with different subcarrier spacings.

3.3 Puncturing concept for URLLC scheduling

The tiling concept is not sufficient, alone, for meeting the requirements for URLLC services, especially for applications with a very bursty traffic. Instead of over-provisioning the URLLC tile, the puncturing (or preemption) mechanism has been proposed in [7] with the following implementation:

- In the downlink, as the scheduling is performed by the base station, if an urgent URLLC packet arrives while all resources are occupied by ongoing eMBB transmissions, the base station preempts some already allocated

resources, leading to a loss for the preempted eMBB user, recovered by retransmissions. This is illustrated in Figure 3 where the third tile is extended in the frequency domain during one slot to serve an urgent URLLC packet.

- In the uplink, when a device has an urgent packet to transmit, it transmits it with a higher power on a resource that may be occupied by an eMBB user, so that the receiver can decode its transmission. For the preempted eMBB packet, it can be recovered iteratively (after decoding the URLLC packet), or by retransmissions.

3.4 Slice identification and mapping

An important issue is the identification of the slice to which a flow of packets belongs, so that a RAN network function needs may apply a potential specific treatment to them. We give the example of the scheduler that is nowadays able to handle a limited number of traffic classes, e.g., via the 4G QoS Class Identifier (QCI). However, the multiplication of slices with specific SLAs will make the task harder for these network functions if they have to handle directly the slice identifier. A RAN slice management entity, such as the Software-Defined Mobile Network Controller (SDM-X) in [5], could facilitate this task by:

(1) performing the function chaining for flows, depending on the SLAs associated to their slices. For instance, Packet Data Convergence Protocol (PDCP) functions related to header compression may be suppressed for some URLLC and mMTC slices [8],
(2) selecting the corresponding configurations at PHY layer for packets belonging to flows of a given slice. For instance, massive MIMO techniques are to be activated for eMBB slices only and channel coding is to be selected based on the reliability targets,
(3) mapping the slice SLA to a QoS identifier and to a tile in order to reduce scheduler complexity.

3.5 Architectural considerations

The network architecture will have a paramount impact on the service latency and the application should be placed closer to the users with critical latency needs. Figure 4 presents three different network architectures which will respond to the different service latency requirements. As presented in (a), when service latency requirement is not very stringent i.e., more than 20 ms, the application can be placed in a centralized application entity. When the application requires an end-to-end latency lower than 10 ms, the application should be lowered to a Centralized Unit (CU), which is also known as Mobile-Edge Computing (MEC). Finally, when the provided service requires very low latency, i.e., 1 ms, the application should be placed in the radio site, the closest possible to the user. A slice manager will have to be integrated at each level i.e., centralized application entity, CU and Radio site, to correctly conduct the traffic corresponding at each slice in each level. Note that, as different slices coexist in the same geographical area, the RAN architecture has to be flexible, allowing the coexistence of the three architectures, with an adequate dimensioning of processing resources at the distributed and centralized units.

3.6 Slice-specific configuration

A slice specific configuration can be performed so that the QoS requirements of the underlying service can be met, including selection of lower layer RAN functions [8]. Table 1 illustrates an example of four slices sharing the infrastructure in a vehicular environment, defined as proposed in section II (i.e., service/industry/customer based):

- Two slices belonging to the same automotive customer, one for entertainment on board (eMBB) and one for safety messages (URLLC),
- One slice for gathering the information sent periodically from sensors on the road (mMTC)
- One eMBB slice for smartphones of pedestrians.

The table reuses some of the concepts developed in the sections above and shows the best per-slice configuration for different network functions.

4 Challenges for RAN slicing

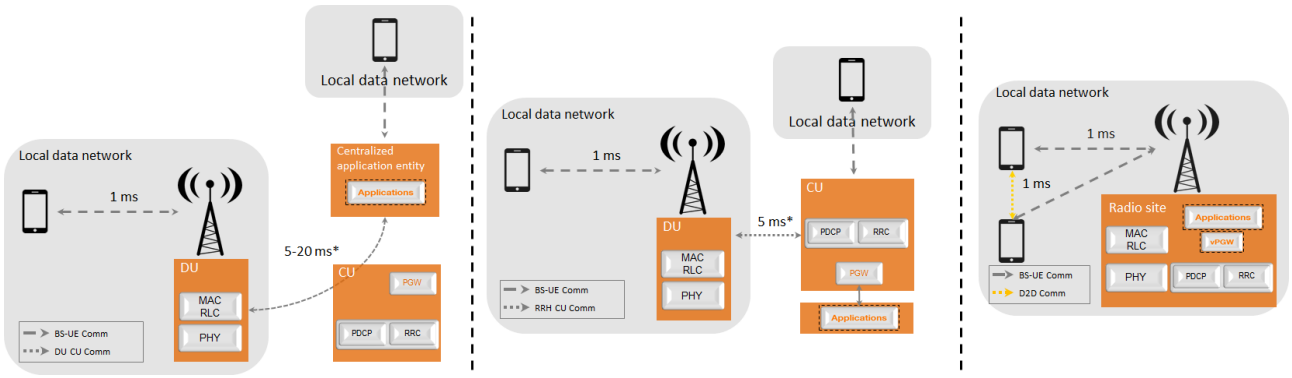
While the above detailed radio features facilitate the implementation of slicing, there are still many challenges when it comes to resource allocation and management.

4.1 Resource allocation

The cohabitation of a large number of slices poses many challenges related to resource allocation to slices and flows, considering not only radio resources but also processing ones (for MEC and virtual radio functions). One of the major open research problems consists in adapting existing resource allocation schemes, designing and developing new ones and fitting them into the new context of slicing in the RAN. These schemes have to ensure jointly QoS for individual slices, fairness between slices and overall resource efficiency. A number of papers provide an overview on fairness in multi-resource allocation. One of the most promising works in this field is that of [13] that proposes the so-called Dominant Resource Fair (DRF) queuing as a scheduling algorithm, generalizing the concept of virtual time for resource sharing in clusters or routers equipped with middle boxes. [14] examines the objectives of mostly advocated resource allocation principles, such as Proportional Fairness (PF), DRF and the proposed Bottleneck Max Fairness (BMF).

While many of these approaches may be relevant in the context of slicing, there are several related open questions:

(1) How fairness is defined for services that do not require high throughputs but rather massive connectivity, high reliability or low latency?
(2) Is a scheduler based on weighting the different slices corresponding to their priorities able to manage resource allocation for the ensemble of traffic of the different tenants served by different slices?
(3) During high load, is puncturing (i.e., preemption of resources by a higher priority service, see 3.3) enough to manage the traffic and service mix?
(4) Is it needed to reserve resources to enforce SLAs for all slices?
(5) If resource reservation is necessary, how much should be reserved and what about over-provisioning?



(a) Wide area coverage slice with distant applications (e.g., eMBB), while the Packet Gateway (PGW) is in the CU.

(b) Slice with regional footprint and medium sensitivity to latency (e.g., augmented reality). Both applications and PGW are in the CU.

(c) URLLC slice with very stringent latency constraints (e.g. industrial control). Virtual PGW (vPGW) and applications are in the radio site.

Fig. 4. Function placement for different slices, focusing on PHY, MAC, Radio Link Control (RLC) and Radio Resource Control (RRC).

Keeping in mind that due to the random nature of traffic, for given reservation levels and scheduling strategy, the answer to the dimensioning exercise is probabilistic. If reservation is performed, the non-reserved resources should be shared among the slices according to the scheduling policy.

In order to be more specific, we discuss the different options for allocating resources to URLLC services. Table 2 presents different URLLC use cases and their generated traffic types and shows how resources have to be allocated to the corresponding slices and how they interact with the general eMBB slice. This table is not meant to be exhaustive, but shows a wide range of resource allocation schemes for these types of services, leading to interesting research perspectives.

4.2 SLA monitoring

Monitoring capabilities are necessary for the customer to verify that the network delivers the desired service with the associated SLA, but it has also the function of alerting the customer in case the SLA cannot be fulfilled anymore. This is essential for critical applications, so that the customer can adopt appropriate security measures, e.g., stop the equipment relying on the network for connectivity, or inform that some functions are temporarily deactivated. Note that the alerting function would take advantage of associating the SLA monitoring to prediction capabilities regarding the risk of SLA non-fulfilment in the future.

Monitoring of legacy RAN is performed using (aggregated) metrics coming from base stations, namely at cell level, or from traces from UEs. The information reported at the radio access level corresponds to the service class granularity (e.g., per QCI in 4G), and hence is less precise than the service granularity at higher layers. As stated earlier, the RAN should be slice aware, and should be able to differentiate slices in order to fulfill the corresponding SLAs. Hence it should be possible to monitor separately users belonging to different slices. In addition, monitoring SLA for slices in a virtual network requires monitoring both physical and virtual resources. We finally note that slice monitoring has to report Key Performance Indictors (KPI) that are user-related

TABLE 1
NETWORK FUNCTION CONFIGURATION FOR DIFFERENT SLICES

	Slice 1 (automotive, eMBB): Entertainment on Board	**Slice 2 (automotive, URLLC): Assisted driving and safety**	**Slice 3 (mMTC): Road sensors**	**Slice 4 (public, eMBB): Smartphones for pedestrians**
Communication mode	UE to Base Station (BS) communications (uplink and downlink)	Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I), exploiting the Device to Device (D2D) communications concepts.	UE to BS communications (uplink)	UE to BS communications (uplink and downlink)
Waveform	OFDM	Filtered-OFDM for coping with the possible asynchronicity [7]	OFDM	OFDM
Subcarrier spacing	30 KHZ for coping with Doppler effect [7]	30 KHZ for coping with Doppler effect	3.75 KHz	15 KHz
MIMO scheme	Predictor antenna for solving the channel aging problem [9]	2*2 Transmit Diversity	Receive diversity	Grid of Beams [10]
Retransmissions	Classical HARQ (Hybrid Automatic Repeat Request)	No HARQ or automatic retransmission (without waiting for acknowledgement) [11]	Multiple transmissions of each packet	Classical HARQ
Scheduling scheme	Proportional fair scheduling	Reserved pool, non-scheduled (contention-based) [11]	Scheduled on multiple uplink resources for providing diversity against collisions	Proportional Fair scheduling
Architecture	Content on a distant server or regional caches	Application server deployed close to the edge, e.g., in road-side units	Centralized data gathering and processing server	Contents on a distant server

(throughputs, delays, reliability), but also network-related (per slice load and energy efficiency). The latter KPIs are challenging to assess as they are related to the infrastructure that is shared between slices.

4.3 Multi-tenant slicing in the RAN

Slicing is expected to facilitate openness of the network to third parties, or tenants (e.g., a virtual network operator or a vertical). Openness of the network to third parties is regarded as a key issue as it may encourage some vertical players to use operators' networks rather than deploying their own infrastructure. At the RAN level, different options for openness are possible, following the degree of control the tenant has on the slice, from monitoring only to full control.

The option that corresponds to the lowest opening degree grants full control to the operator, while the third party is allowed to access to monitoring KPIs. In this case, the operator is in charge of guaranteeing the fulfillment of the SLA, that corresponds to e.g., the QoS provided to the end-users of the slice. The operator is the owner of the network and is responsible for the network operation and management.

The other extreme option corresponds to a full control of the network slice by the third party. The operator only provides the infrastructure to the third party which is in charge of operating the slice. Here we may have several sub-options: the slices are pre-designed by the operator and are provided to the third party as "plug-and-play" slices; or the third party performs design/composition of the slice by using virtual functions offered by the operator or even by onboarding its own functions. From a RAN point of view, as the operator is leasing the control of a part of its network to a third party, the slice should be deployed on a dedicated spectrum, which may raise regulatory issues if the slice is deployed on licensed spectrum.

In the intermediate option, the slice is operated by the MNO and partially controlled by the third party. The slice owner can control some network functions or change some configurations in the network. Consider for example the "Factory of the Future" vertical, where the slice is typically deployed on a limited area, and where the slice owner may want to have a higher involvement in the slice design and the network optimization functions. For instance, the throughput/coverage required in each area of the factory may change in time depending on mobile robots' trajectories, installation of new machines, changes in the internal organization targets, etc. The slice owner may want to re-plan the resources accordingly, including transmitted powers and beamforming schemes. This option is feasible as long as the isolation of the slice is guaranteed; in particular, the actions performed on this slice do not harm the rest of the operator network. This can be performed for example through the selection of certain network functions and their allowed configurations (configurable parameters and their ranges).

More generally, this intermediate openness raises several issues about the impact of a specific third party on the global performance of the network, and on defining the responsibilities of the slice owner and of the operator when performance degradation occurs in the network.

TABLE 2

RESOURCE ALLOCATION PRINCIPLES FOR URLLC SLICES AND THEIR INTERACTIONS WITH THE EMBB SLICE. DETAILED USE CASE DESCRIPTIONS [2].

Use case	Traffic type	Resource allocation for the corresponding URLLC slice	Interaction with eMBB slice
Industry automation (controlled indoor environment)	Periodic generation of packets by a limited number of machines	Persistent (cyclic) resource reservation for each UE [12]. As the amount of resources is to be pre-determined and cannot be changed dynamically, an over-allocation is advocated along with a continuous channel estimation that adapts the allocation to radio condition variation on a scale of several milliseconds.	Any resources in the URLLC traffic cycle that are not reserved for URLLC can be used for eMBB. The amount of reserved resources is to be updated depending on variations in traffic and radio conditions
	Sporadic generation of packets by a large number of machines	Cyclic reservation of a pool of common resources. Several copies of each packet may be sent for combating collisions and different copies may be combined for increasing reliability [11].	
Traffic safety (vehicular environment)	Periodic packets sent by cars	Reservation of a pool of common resources.	eMBB and URLLC resources are orthogonal in the same cells, but interference may happen between URLLC service in one cell and eMBB service in another cell that, for instance did not activate the URLLC slice.
	Correlated packet generation due to unexpected events	Preemption of resources from the other slices when the packet flood is detected	eMBB resources are pre-empted for serving high priority URLLC traffic. Higher-layer actions are needed, e.g., change in video encoding, otherwise dropping may occur.
Tactile Internet (wide area coverage)	Persistent traffic generated by well-localized UEs (e.g., around medical centers)	URLLC slice is always on and resources are continuously reserved on the end-to-end path	eMBB and URLLC resources are orthogonal in the same cells, but interference may happen with other cells that do not have the URLLC slice activated
	Occasional point-to-point slice establishment (e.g., for emergencies)	Ad-hoc URLLC slice establishment and resource reservation	Some eMBB QoS problems may occur at URLLC slice establishment. The eMBB slice is to be reconfigured (e.g., via admission control, dropping or higher layer mechanisms) as tactile Internet slices are in general long-living.

7 CONCLUSION

This paper has investigated the slicing concept in the 5G RAN with the objective of identifying the plausible options for implementing the slicing concept at the RAN level to respond to the needs of verticals. While the lower layers need not know the exact ID of the slice, they should be able to respond to their heterogeneous requirements. Different concepts and challenges related to RAN slicing and their management have been described. The tiling concept allows implementing different PHY/MAC configurations on the same spectrum, and a slice management entity maps slices to tiles at a relatively long time scale. The low level scheduler that operates per-tile dynamically allocates resources to flows while the differentiated scheduling mechanism remains an open research problem. The degree of openness of the RAN slices to third parties or tenants, such as verticals, that may be willing to be involved in slice monitoring and configuration, in another open issue that introduces new challenges to RAN management.

REFERENCES

[1] S. Elayoubi (ed.), "5G innovations for new business opportunities", 5G PPP white paper for the Mobile World Congress, March 2017.

[2] S. Elayoubi and M. Maternia (eds.), "5G-PPP use cases and performance evaluation modeling", 5G PPP white paper, April 2016.

[3] X. Foukas *et. al.,* "Network Slicing in 5G: Survey and Challenges", IEEE Communications Magazine, 55(5), 94-100, 2017.

[4] 3GPP TS 23.501 V1.3.0 (2017-09) Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15).

[5] P. Rost *et. al.,* "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks", *IEEE Communications Magazine*, 55(5), 72-79, May 2017.

[6] K. I. Pedersen *et. al.*, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53-59, March 2016.

[7] 5G PPP FANTASTIC-5G project, Deliverable D3.2, "holistic link solution adaptation", April 2017.

[8] I. da Silva *et. al.*, "Impact of network slicing on 5G Radio Access Networks.", *IEEE Networks and Communications (EuCNC), 2016.*

[9] D. T. Phan-Huy *et. al.* , "5G on Board: How Many Antennas Do We Need on Connected Cars?," *IEEE Globecom Workshops*, Washington, DC, 2016, pp. 1-7.

[10] B. Panzner *et. al.*, "Deployment and implementation strategies for massive MIMO in 5G," *IEEE Globecom Workshops*, Dec. 2014.

[11] B. Singh *et. al.*, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions." *IEEE Wireless Communications Letters, 7(2), 182-185, 2018*.

[12] Y. Han, S. E. Elayoubi, A. Galindo-Serrano, V. S. Varma, and M. Messai, "Periodic Radio Resource Allocation to Meet Latency and Reliability Requirements in 5G Networks," *IEEE VTC2018-Spring*, June 2018.

[13] Ali Ghodsi *et. al.*, "Multi-resource fair queueing for packet processing". ACM SIGCOMM 2012, New York, NY, USA, 1-12.

[14] Thomas Bonald and James Roberts, "Multi-Resource Fairness: Objectives, Algorithms and Performance". *ACM SIGMETRICS 2015*, New York, NY, USA, 31-42.

Salah Eddine Elayoubi received the M.S. degree in telecommunications from the National Polytechnic Institute of Toulouse, France, in 2001, and the Ph.D. and Habilitation degrees in computer science from the University of Paris VI, Paris, France, in 2004 and 2009, respectively. From 2004 to 2013 he was with Orange Labs in France. Since January 2018, he is associate professor at CentraleSupélec, France. His research interests include radio resource management, modeling, and performance evaluation of mobile networks.

Sana Ben Jemaa received the engineering diploma of the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 2001 and her Ph.D. degree in electrical engineering in the École Nationale Supérieure des Télécommunications, Paris, France in 2004. Since 2004 she has been working as a research engineer and project manager at Orange Labs. Her research interests include mobile/wireless communications, self-organising networks and artificial intelligence application for RAN management.

Zwi Altman received the B.Sc. and M.Sc. in the Technion-Israel Institute of Technology, in 1986 and 1989 respectively, and the Ph.D. from the INPT France in 1994. From 1994 to 1996 he was a Post-Doctoral Research Fellow in the University of Illinois at Urbana Champaign. He joined Orange Labs in 1996 where he has been involved in various projects on self-organizing networks, autonomics, SDN, network slicing, massive MIMO and quantum communication. Dr. Altman is a senior research expert in Orange.

Ana Galindo-Serrano received her telecommunication and electronic engineering degree from Instituto Superior Politcnico Jos Antonio Echeverria, La Habana, Cuba, in 2007 and the PhD at the Dept. of Signal Theory and Communications (TSC) of the Technical University of Catalonia (UPC) in 2012. In 2012 she joined Orange Labs as a research engineer. Her research topics of interest interest are mobile communication sytems design, energy savings and sustainable development.