ACCEPTED FROM OPEN CALL

On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration

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

Through a comprehensive analysis of the impact that the realization of RAN slicing has on the different layers of the radio interface protocol architecture, the authors propose a framework for the support and specification of RAN slices based on the definition of a set of configuration descriptors that characterize the features, policies and resources to be put in place across the radio protocol layers of a next generation RAN node.

ABSTRACT

Network slicing is a fundamental capability for future Fifth Generation (5G) networks to facilitate the cost-effective deployment and operation of multiple logical networks over a common physical network infrastructure in a way that each logical network (i.e. network slice) can be customized and dimensioned to best serve the needs of specific applications (e.g. mobile broadband, smart city, connected car, public safety, fixed wireless access) and users (e.g. general public, enterprise customers, virtual operators, content providers). The practical realization of such capability still raises numerous technical challenges, both in the Core and RAN parts of the 5G system. Through a comprehensive analysis of the impact that the realization of RAN slicing has on the different layers of the radio interface protocol architecture, this article proposes a framework for the support and specification of RAN slices based on the definition of a set of configuration descriptors that characterize the features, policies and resources to be put in place across the radio protocol layers of a next-generation RAN node.

INTRODUCTION

Unlike current 4G systems mainly designed to provide a "one size fits all" mobile broadband solution, 5G systems are intended to simultaneously support a wider range of application scenarios and business models (e.g. automotive, utilities, smart cities, high-tech manufacturing) [1]. This expected versatility comes with a high variety of requirements on network functionalities (e.g. security, mobility, policy control features) and expected performance (e.g. peak rates above 10Gb/s, latencies below 1 ms with 10^{-5} reliability, 500 km/h mobility target) that cannot always be met through a common network setting (e.g. optimizing the network for low latency with high reliability could come at the expenses of reduced spectral efficiency). In this context, support for network slicing in 5G systems has become a foundational requirement to allow 5G system operators to compose and manage dedicated logical networks with specific functionality, without losing the economies of scale of a common infrastructure [2]. Each of these logical networks, referred to as *network slices*, can be tailored to fulfil at least a couple of purposes:

•To provide a particular system behavior (i.e. slice type) through the use of specific control plane (CP) and/or user plane (UP) functions to best support specific service/application domains (e.g. optimized protocols for enhanced MBB [eMBB], massive Machine Type Communications [mMTC], Ultra-Reliable and Low Latency Communications [URLLC]). For instance, a user equipment (UE) for smart metering applications can be served through a network slice with radio access tailored to very small, infrequent messages and with no need to implement unnecessary functions (e.g. no mobility support).

•To provide a particular tenant (i.e. an organization or business entity entitled to use the network slice) with a given level of guaranteed network resources and isolation with regard to the operation of other concurrent slices. For instance, UEs/subscribers of a public safety (PS) agency can be served through a network slice that guarantees a minimum capacity during network congestion periods.

3GPP has recently completed the normative specifications regarding service and operational requirements to support network slicing [3], and work has started on both system architecture aspects [4] and related management and orchestration capabilities [5]. Simultaneously, the network slicing concept is being addressed in the 5G architectures currently under development in different research projects such as 5G-NORMA [6], METIS-II [7] or SESAME [8]. Indeed, a complete solution for network slicing combines multiple facets, ranging from virtualization techniques for the abstraction and sharing of radio resources (e.g. network virtualization substrate concept in [9]) to network slice lifecycle management solutions enabling the delivery of *Network Slice as a Service* (e.g. 5G network slice broker concept in [10]).

In this context, this article elaborates on the realization of network slicing within the Radio Access Network (RAN), which still poses multiple open questions. To this end, the article first tackles the overall network slicing architectural framework for 5G systems and identifies the fundamental design challenges for the realization of RAN slicing. Then, a solution approach is proposed through the definition of a set of descriptors to parametrize the features, policies and resources within the radio protocol layers for the configuration of a RAN slice. The systematic and comprehensive analysis conducted is, to the authors' best knowledge, the first attempt to tangibly structure and specify how a RAN slice can be realized and how various RAN slices can be

Digital Object Identifier:
10.1109/MCOM.2017.1700268

The authors are with the Universitat Politècnica de Catalunya (UPC).

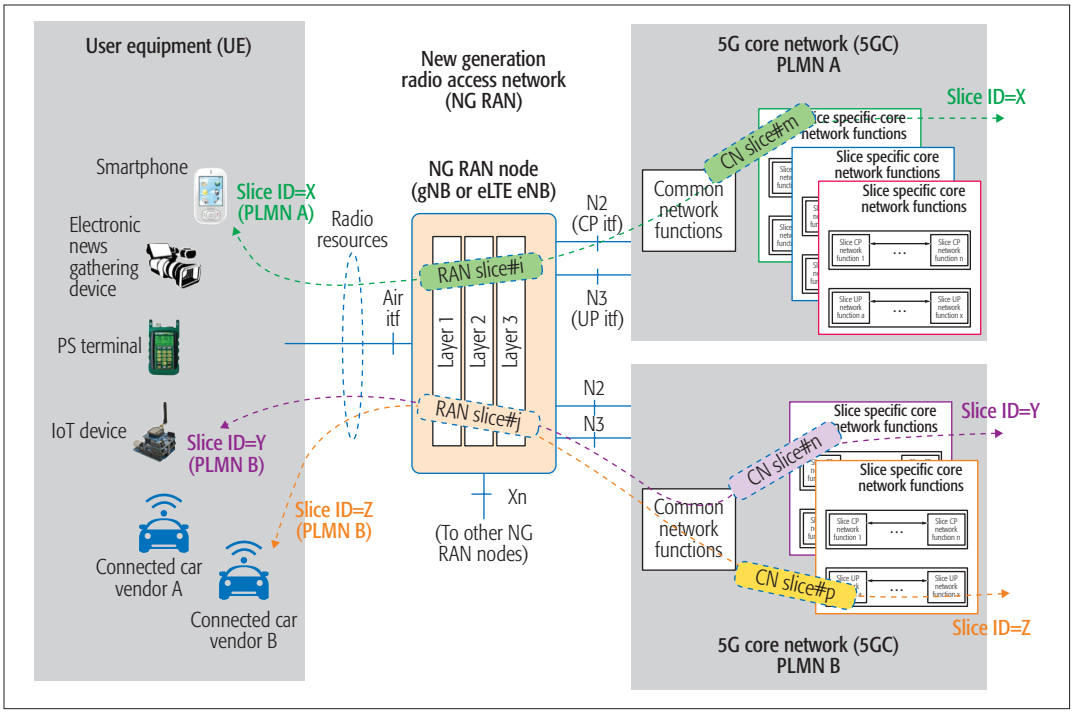

Figure 1. Illustration of the network slicing architecture.

separately customized. Finally, the applicability of the proposed RAN slicing solution framework is illustrated.

Architectural Framework and RAN Slicing Design Challenges

The realization of network slices considers, in the most general case, support for specific features and resources both in the 5G Core Network (5GC) part, which we will refer to as a CN slice, and in the New Generation RAN (NG-RAN) part, referred to as a RAN slice. A network slice is to be uniquely identified by a *Slice_ID* within a 5G network, the latter identified by a Public Land Mobile Network (PLMN) identity. The *Slice_ID* could take standard values to facilitate slicing configurations across networks in roaming scenarios or just remain PLMN-specific [4].

On the 5GC side, CN slices are to be realized through the deployment of a combination of a highly modularized set of 5GC functionality for CP and UP functions, including network functions (NFs) for e.g. network registration and mobility management (i.e. access and mobility management function [AMF]), 5G connectivity service handling (i.e. session management function [SMF]), user plane forwarding and QoS handling (i.e. user plane function [UPF]), and so on. As illustrated in Fig. 1, some NF instances could be common to several CN slices, that is, supporting and logically belonging to more than one slice, while other NF instances just serve specific CN slices (e.g. the AMF instance, which conveys all the UE-5GC signaling for registration area management, UE reachability in idle state, and so on, has to be the same instance for all the network slices serving a given UE [4]). All these 5GC NFs are likely to be implemented as virtualized network functions (VNFs) running on multi-tenant cloud infrastructures [11] and flexibly orchestrated as required.

On the other hand, the NG-RAN fundamentally consists of gNBs and/or eLTE eNBs, which are single NFs that provide the UP/CP protocol terminations toward the UEs and embed all the radio access functionality. Specifically, a gNB is a NG-RAN node operating the new radio (NR) interface while an eLTE eNB is conceived as the evolution of the legacy eNB to support connectivity to both legacy CN (i.e. Evolved Packet Core [EPC]) and 5GC and so facilitate co-existence and migration options [12]. As illustrated in Fig. 1, a quite diverse range of UEs could be connected to the same NG-RAN node, though potentially served via different network slices. In this context, NFV technologies can also play a role in the realization of RAN slicing as long as a part of the RAN node functionality can be deployable as a VNF (e.g. a gNB node functionally split between a centralized unit, deployed with NFV, and a distributed unit with the specific RF components). However, regardless of the deployment options, the realization of the RAN slices requires addressing how the pool of radio resources (i.e. RF bandwidth) allocated to one NG RAN node can be configured and operated to simultaneously deliver multiple and diverse RAN behaviors, turning RAN slicing support into a much more challenging issue. In this respect, let us consider that the RF bandwidth operated by a NG-RAN node can be flexibly arranged into a number of RF channels with diverse transmission bandwidths. A simple approach to support multiple RAN slices with slice-specific/optimized waveforms would be to implement each slice type on different cells using separate RF channels and radio protocol layer instances (i.e. Layer 1, 2 and 3 functions), similarly as done today with the multiple radio access technologies (RATs) that co-exist in the current RAN infrastructures (e.g. UMTS cells, LTE cells, NB-IoT cells). However, scalability in terms of the number of supported slice types and reduced resource

Unlike current 4G systems mainly designed to provide a "one size fits all" mobile broadband solution, 5G systems are intended to simultaneously support a wider range of application scenarios and business models (e.g. automotive, utilities, smart cities, high-tech manufacturing)

From a functional perspective, it is proposed to specify the operation of each RAN slice through a set of configuration descriptors that parametrizes the features, policies and resources put in place across the radio protocol layers of the RAN node.

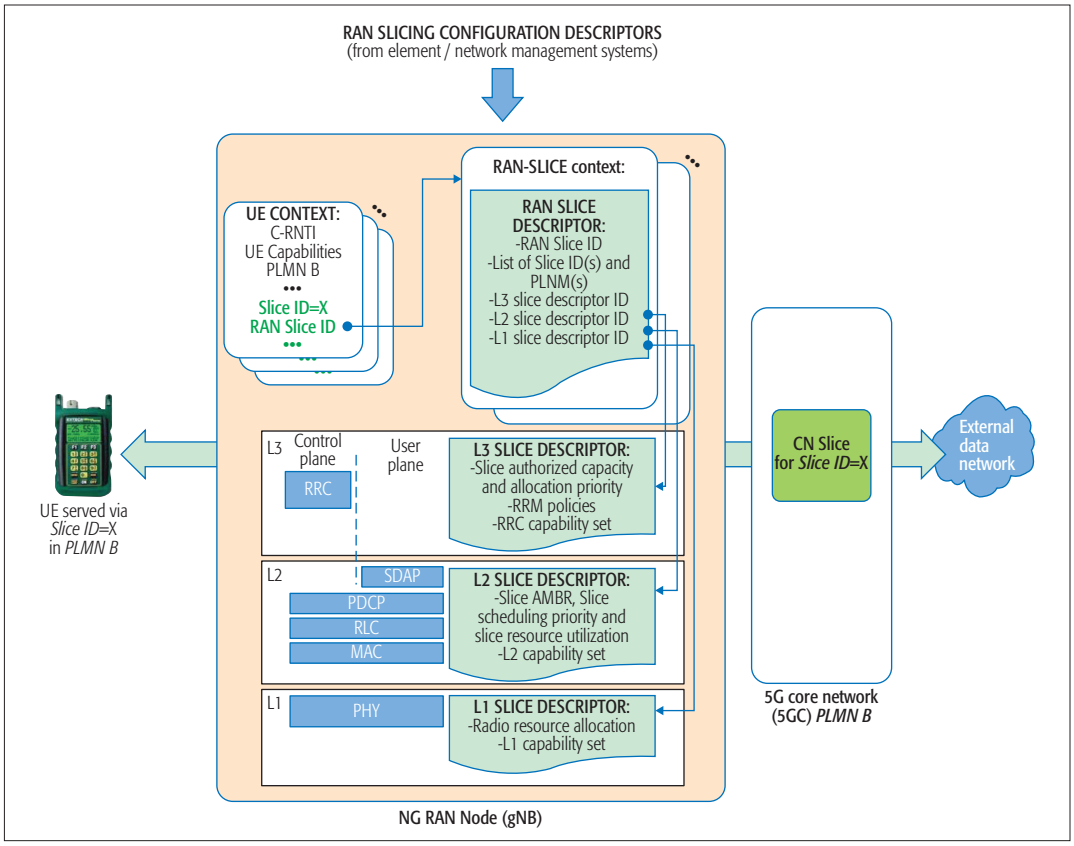

Figure 2. Framework for the realization of RAN slices in a NG-RAN node.

efficiency (i.e. no traffic multiplexing gains among slices) would be clearly major drawbacks. Therefore, the fundamental design challenge strives for the realization of multiple RAN slices that can be concurrently multiplexed in a single cell while at the same time achieving an efficient use of the radio resources. This entails tackling the following open questions:

- How to imbricate a number of slice-specific radio interface protocols over the same cell.
- How to manage radio resource allocation to UEs within a cell so that pre-established levels on capacity and isolation can be offered per RAN slice.
- How to enable support for optimized radio resource management (RRM) configurations and policies on a per-slice basis (e.g. admission policies, mobility control).

Realization of RAN Slices

Taking as a starting point the initial work conducted within 3GPP on NG-RAN and NR [12, 13] as well as the radio interface protocol architecture already consolidated in current 3G/4G RANs, the analysis presented here shows that a set of new blocks of information, configuration descriptors and protocol features has to be introduced across the protocol layers of a NG-RAN node in order to address the mentioned open questions. In this respect, the proposed overall framework for RAN slicing support within a NG-RAN node is illustrated in Fig. 2 and explained in the following.

From an operational perspective, a *UE Context* is instantiated within the RAN at the time the UE becomes active and establishes a logical control connection (i.e. radio resource control [RRC] connection). The *UE Context*, populated from information exchanged between the RAN and both the UE and CN, is a block of information that contains all the necessary data required to maintain the RAN services toward the UE (e.g., temporary identifiers such as the cell radio network temporary identity [C-RNTI], security context, UE capabilities, QoS related information, etc.). On this basis, we propose to add a RAN slice identifier (*RAN_Slice_ID*) to the *UE Context* and use it as a pointer to a new block of information within the NG-RAN node, denoted as RAN Slice Context, containing all the data necessary to support the operation of a particular RAN slice along with the *Slice_ID(s)* that are served through the RAN slice. The selection of the *RAN_Slice_ID(s)* is to be conducted by the NG-RAN node based on the *Slice_ID(s)* signaled by the UEs during the initial attach procedure or indicated from CN in subsequent signaling (the specific signaling procedures are still a work in progress within the 3GPP [4]).

Accordingly, from a functional perspective, it is proposed to specify the operation of each RAN slice through a set of configuration descriptors that parameterize the features, policies and resources put in place across the radio protocol layers of the RAN node. In particular, a *RAN Slice Descriptor* is introduced as the baseline descriptor for the instantiation of a *RAN Slice Context* within a RAN node. As illustrated in Fig. 2, the *RAN Slice Descriptor* includes at least the *RAN_Slice_ID*, the list of associated *Slice_ID(s)* and *PLMN_ID(s)*, and a set of pointers to the configuration descriptors of the underlying radio protocol layers 3, 2 and 1 (L3, L2, L1) for the realization of the RAN slice.

L3 Slice Descriptor: L3 comprises the RRC

protocol and RRM functions such as radio bearer control (RBC), radio admission control (RAC) and connection mobility control (CMC) for the activation and maintenance of radio bearers (RB), which are the data transfer services delivered by the radio protocol stack. For each UE, one or more user plane RBs, denoted as data RBs (DRBs), can be established per protocol data unit (PDU) session, which defines the connectivity service provided by 5GC [12]. It is worth noting that traffic flows from different slices are not served by the same DRBs since traffic for different slices is handled through different PDU sessions [4]. Accordingly, a *L3 slice descriptor* is necessary to specify the capacity allocation for the RAN slice (e.g. number and characteristics of the DRBs that can be simultaneously established), the RRM policies that govern the operation of the slice (e.g. DRB configuration policies) and the capability set of the RRC protocol in use (e.g. application type specific RRC messages).

L2 Slice Descriptor: L2 comprises a Medium Access Control (MAC) sub-layer for multiplexing and scheduling the packet transmissions of the DRBs over a set of transport channels exposed by L1. Moreover, L2 embeds a number of processing functions configurable on a per-DRB basis for e.g. segmentation, Automatic Repeat reQuest (ARQ) retransmissions, compression and ciphering (i.e. Radio Link Control [RLC] and Packet Data Convergence Protocol [PDCP]). In the NR specifications, an additional L2 sub-layer named the Service Data Adaptation Protocol (SDAP) is included to map the DRBs and the traffic flows managed by the 5GC, referred to as QoS flows [12]. Therefore, considering that the current MAC operation is based on individual UE and DRB -specific QoS profiles, a *L2 slice descriptor* is necessary to define the packet scheduling behaviors to be enforced on the traffic aggregate of DRBs of the same slice and to specify the capability set of the applicable L2 sub-layers processing functions.

L1 Slice Descriptor: L1 provides L2 with transfer services in the form of transport channels, which define how the data is transferred (e.g. transmission time interval [TTI], channel coding). L1 also establishes the corresponding radio resource structure of the cell radio resources (e.g. waveform characteristics and time-/frequency-domain resource structure). Considering that a RAN slice may require specific L1 transfer service capabilities (e.g. low latency shared transport channel) and/or specific radio resource allocation of the cell radio resources, a *L1 slice descriptor* is needed to specify both aspects.

More details on the above introduced descriptors and their impact on the configuration and extended features necessary for RAN slicing within L3, L2 and L1 are discussed below.

L3 Configuration

When multiple RAN slices are realized over shared radio resources, the RRM functions for RBC, RAC and CMC have to assure that each RAN slice gets the expected amount of resources and, in case, handles any resource conflicts that might appear across slices. These RRM functions are typically implemented as vendor-specific algorithms only abided by the use of the data models and control signaling capabilities established in the standards (e.g. UE and RB QoS model parameters, RRC protocol messages and procedures). Therefore, we propose to specify the following set of parameters per RAN slice to dictate the operation of the RRM functions for capacity allocation and traffic isolation among the slices:

•*Slice Authorized Capacity:* This can be a combination of resource-oriented and rate-oriented parameters that limit the number and characteristics of the RBs established for the entire slice. Resource-oriented parameters can include absolute or relative occupation levels of the consumed radio resources (i.e. radio resource limitations) as well as hard limits on the number of simultaneously established *UE Contexts*/RBs (i.e. license limitations). Rate-oriented parameters can include rate limits on the aggregate bit rate of the entire set of admitted guaranteed bit rate (GBR) RBs within the slice. All these slice capacity parameters are to be used by the RAC for the admission/rejection of RBs as well as by the RBC/CMC functions in order to decide on modifications, handovers or even forced releases of active RBs if network dynamics turns into increased radio resource consumption in excess of the established limits.

•*Slice Allocation Priority:* This parameter allows for conflict resolution among UE/RB resource requirements across slices that cannot be solved based only on the *Slice Authorized Capacity* parameters nor using the policies associated with the individual UE/RBs (e.g. a situation in which there are two GBR DRB admission requests with the same QoS profile in distinct RAN slices that cannot be granted simultaneously due to temporary congestion). In legacy 3G/4G RANs, priority and pre-emption policies at the UE/RB level are solved through the allocation and retention priority (ARP) parameter included in the QoS profile. The ARP encodes information about priority level (scalar with 15 levels), pre-emption capability (flag with "yes" or "no") and pre-emption vulnerability (flag "yes" or "no"). A similar semantic can be adopted for the *Slice Allocation Priority* parameter.

Moreover, since pursuing different optimization targets for the efficient use of the radio resources requires RRM functions with different parameterization, the RRM policies should be specifiable on a per-slice basis. This can be accomplished through a prescriptive specification, establishing the RRM algorithms and associated configuration parameters (e.g. thresholds, timers), and/or a declarative specification, establishing the expected RRM behavior (e.g. key performance indicators [KPIs] and optimization goals).

Regarding the RRC protocol, the customization of RRC messages and procedures on a per-slice basis (e.g. specific RRC signaling for Narrow Band Internet of Things [NB-IoT] applications) can be realized without any impact on the protocol stack if customized RRC signaling is only used through UE-dedicated signaling RBs (SRBs) and/or through common logical channels not shared with other RAN slices. However, if multiple RAN slices are configured to share the same set of common logical channels (e.g. broadcast control channel [BCCH], paging control channel [PCCH] and common control channels [CCCH]), the following extended features have to be incorporated within the RRC protocol:

When multiple RAN slices are realized over shared radio resources, the RRM functions for RBC, RAC and CMC have to assure that each RAN slice gets the expected amount of resources and, in case, handles any resource conflicts that might appear across slices.

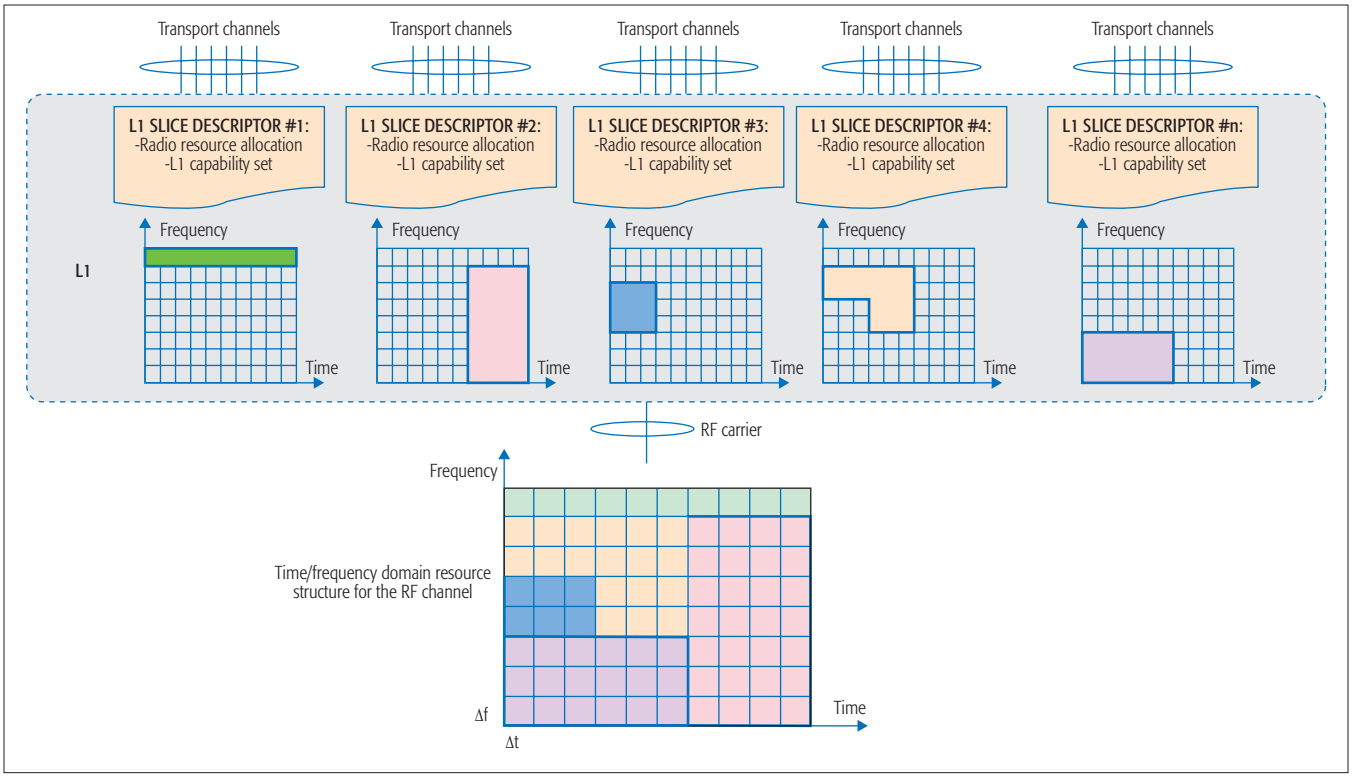

Figure 3. Illustration of slicing support in L1.

- Protocol fields within the RRC messages to allow UEs to discriminate among signaling from different slices.
- System information block (SIB) messages to advertise through a common BCCH the *Slice_ID(s)* that can be reached from the cell. This allows the UE to take into account this information for network discovery and selection processes.
- SIB messages to support cell (re-)selection parameters and neighboring cell information broadcasting per *Slice_ID*, so that the behavior of terminals in idle mode can be set differently per-slice.
- SIB messages to support access barring and load control (AB/LC) per *Slice_ID* so that un-scheduled transmissions over the uplink CCCH can be controlled separately per slice.
- Paging configuration features allowing paging cycles to be organized considering the specific needs of each *Slice_ID*.

L2 Configuration

In current 3G/4G systems, radio resource scheduling by the MAC takes account of the traffic volume and radio conditions per UE together with the UE/RB QoS parameters that define the expected forwarding behavior (e.g. QoS class identifier [QCI], GBR, maximum bit rate [MBR], aggregated MBR per UE [UE-AMBR]). While this approach suffices for QoS differentiation on the packet forwarding treatment per-UE/RB, it lacks any formalization to parameterize the expected QoS behavior at the packet forwarding level for the traffic aggregated within a given RAN slice.

Therefore, the proposed *L2 slice descriptor* includes the following parameters to dictate, along with the per-UE/RB QoS profiles, the operation of the MAC scheduler and yield isolation at packet forwarding level on a per-slice basis:

- *Slice-AMBR*, to limit the aggregate bit rate of all the non-GBR RBs associated with the slice.
- *Slice Scheduling Priority*, to handle short-term traffic congestion conflicts between RBs with the same QoS profile (e.g. same QCI) but belonging to slices that should be given different precedence treatment.
- *Slice Resource Utilization*, used to establish constraints on the amount of physical-layer resources scheduled by the MAC that are consumed by the slice. This constraint can be formulated as a percentage of the overall L1 resources that are managed by the MAC scheduler.

In addition, the *L2 slice descriptor* includes an *L2 capability* set to establish the possible configuration options for the RBs associated with the slice, including Hybrid ARQ configurations (e.g. synchronous/asynchronous operation and number of processes in parallel), RLC operation modes (e.g. acknowledged/unacknowledged/transparent modes and status reporting) and PDCP options (e.g. ciphering support).

L1 Configuration

The new physical layer for 5G NR [13] is being defined with the goal to provide high flexibility for the use of different waveforms (e.g. orthogonal frequency-division multiplexing [OFDM]-based waveforms with different numerologies) and adaptable time-frequency frame structures (e.g. selectable slot durations, dynamic assignment of DL/UL transmission direction). Considering that the L1 optimal settings can differ per slice type,

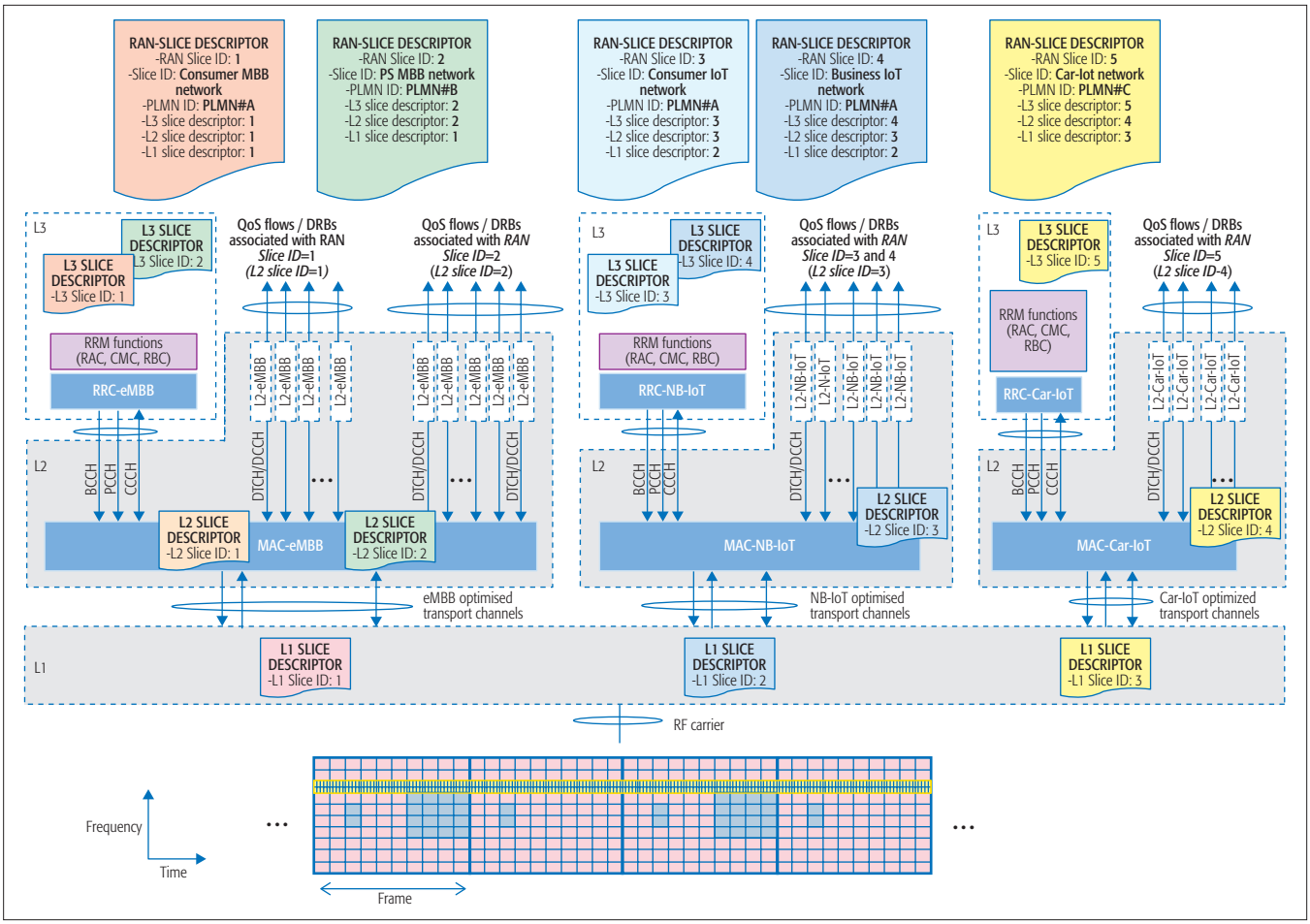

Figure 4. Radio protocol stack view of the RAN slicing configuration.

the proposed L1 descriptor intends to establish a partitioning of the L1 radio resource structure so that different L1 optimization settings can be simultaneously applied.

The proposed approach is illustrated in Fig. 3. Each tile represents the smallest allocation unit of time duration Δt and frequency size Δf (e.g., Δt and Δf could be an integer number of, respectively, OFDM symbols and subcarriers with the minimum supported subcarrier separation). On this basis, the overall RF carrier resources are split off in a number of *L1 slices* that could be separately optimized to offer specific transfer service capabilities (e.g. different TTI sizes, synchronous/asynchronous access). The mixing of L1 slices in the same resource grid with different waveforms characteristics could be achieved through the use of the different OFDM numerologies and transmitter windowing techniques that mitigate the potential inter-slice interference [14]. Moreover, this segmentation of resources per slice avoids the transmission of common reference and control signals over the entire RF bandwidth, but instead uses self-contained transmissions per slice that can be arranged as needed (e.g. one slice could be configured to use a strict time-division separation of the physical layer control and data, as in current LTE, while in another slice control and data could be multiplexed over the same radio resources [15]). Note that this approach requires a minimum amount of resources to be allocated within the resource grid structure to facilitate the cell search process (e.g. common reference signals for UEs to acquire time and frequency synchronization) and provide the information for UEs to locate the control channels within the L1 slices. This approach also allows the segmentation of resources to be dynamic (e.g. the location and number of resource units allocated to each L1 slice adapted over time to match traffic demand variations).

RAN Slicing Configuration Examples

To gain insight into the proposed framework, let us consider an illustrative scenario for the applicability of RAN slicing involving multiple service providers, service types and groups of subscribers/applications. Let us assume a commercial MNO (MNO#A — PLMN#A) that has deployed a NG-RAN and primarily exploits it to deliver eMBB and IoT services to retail customers. To this end, MNO#A sets up three RAN slices: one for eMBB and two for mMTC. The eMBB slice (*RAN_Slice_ID*=1) is used for general public mobile broadband services, one of the mMTC slices (*RAN_Slice_ID*=3) for small enterprise customers with IoT needs (e.g. transportation companies), and the other mMTC slice (*RAN_Slice_ID*=4) for a few customers with large-scale deployments of IoT applications (e.g. utilities). In addition, MNO#A provides wholesale access to two third party service providers, leasing capacity from its NG-RAN infrastructure. One third party provider is a PS communications operator (MNO#B) that

Parameter		Value
Deployment settings	Path loss model	Urban micro-cell model with hexagonal layout (source: Report ITU-R M.2135 "Guidelines for evaluation of radio interface technologies for IMT-Advanced")
	Shadowing standard deviation	3 dB in LOS and 4 dB in NLOS.
	Base station antenna gain	5 dB
	Frequency	3.6 GHz (source: RSPG16-032 FINAL "Europe Radio Spectrum Policy Group Opinion on spectrum for 5G")
	Transmitted power per PRB	14 dBm
	Carrier bandwidth	100 MHz, equivalent to 500 PRBs for OFDM numerology with f=15kHz
	UE noise figure	9 dB (considered the same for all types of devices)
	Spectral efficiency model to map SINR with bit rate	*L1_Slice_ID*=1: Model A.1 in 3GPP TR 36.942 and assuming a maximum spectral efficiency of 8.8 b/s/Hz *L1_Slice_ID*=2 and 3: 1 b/s/Hz
RAN_Slice_ID=1	Traffic demand	Mix of GBR and non-GBR traffic, modelled with Poisson session arrival time distribution and exponential session duration. GBR DRBs used to offer 10 Mb/s high-quality video streaming services. Average of 100 Mb/s for GBR traffic. Average of 10 simultaneous active DRBs for non-GBR traffic.
	L3 Descriptor – *Slice Authorized Capacity*	Maximum aggregated GBR traffic = 240 Mb/s or up to 56 percent of PRBs consumed on average in *L1_Slice_ID*=1
	L2 Descriptor – *Slice Resource Utilization*	Up to 70 percent of the PRBs left available for non-GBR traffic in L1_Slice_ID=1
	L1 Descriptor – *Radio resource allocation*	450 PRB (shared with *RAN_Slice_ID*=2)
RAN_Slice_ID=2	Traffic demand	Mix of GBR and non-GBR traffic, modelled with Poisson session arrival time distribution and exponential session duration. GBR DRBs used for MCVideo with 2 Mb/s standard video quality. Average of 30 Mb/s for GBR traffic. Average of five simultaneous active DRBs for Non-GBR traffic.
	L3 Descriptor – Slice Authorized Capacity	Maximum aggregated GBR = 60 Mb/s or up to 14 percent of PRBs consumed on average in *L1_Slice_ID*=1
	L2 Descriptor – Slice Resource Utilization	Up to 50 percent of the PRBs left available for non-GBR traffic in L1_Slice_ID=1
	L1 Descriptor – Radio resource allocation	450 PRB (shared with *RAN_Slice_ID*=1)
RAN_Slice_ID=3	Traffic demand	Mix of IoT applications (e.g. assets tracking, environmental monitoring, etc.) whose traffic aggregate is characterized by messages of 128 Bytes every one minute on average. Total number of connected devices is 1000.
	L3 Descriptor – *Slice Authorized Capacity*	Maximum number of UE contexts = 2000
	L1 Descriptor – *Radio resource allocation*	1 PRB (shared with *RAN_Slice_ID*=4)
RAN_Slice_ID=4	Traffic demand	Smart meter application that generates 1024 Bytes messages every two minutes. Total number of connected devices is 400.
	L3 Descriptor – *Slice Authorized Capacity*	Maximum number of UE contexts = 500
	L1 Descriptor – *Radio resource allocation*	1 PRB (shared with *RAN_Slice_ID*=3)
RAN_Slice_ID=5	Traffic demand	Full buffer model
	L1 Descriptor – *Radio resource allocation*	49 PRB

Table 1. Simulation parameters.

contracts an eMBB RAN slice (*RAN_Slice_ID*=2) for mission critical push-to-talk (MCPTT) and MCVideo services in a way that supplements the capacity of its own dedicated PS broadband network (PLMN♯B). The second provider is an IoT communications provider (MNO♯C — PLMN♯C) specialized within the automotive sector, who exploits the leased RAN slice (*RAN_Slice_ID*=5) customized to serve the needs of connected car applications.

Figure 4 provides a detailed illustrative view of the radio interface protocol architecture and the corresponding descriptors for the realization of this scenario. The overall configuration consists of five *RAN Slice Descriptors*, including a combination of five *L3*, four *L2* and three *L1 Slice Descriptors*. At L1, *L1_Slice_ID*=1 is configured for eMBB services (based on e.g. a OFDM waveform as currently used for LTE), *L1_Slice_ID*=2 for mMTC services (based on e.g. the legacy NB-IoT physical layer) and *L1_Slice_ID*=3 for the delivery of URLLC services (based on e.g. a new 5G waveform optimized for vehicle-to-infrastructure communications, denoted as Car IoT in this example).

			Time instants								
			t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
Traffic demand variation on *RAN_Slice_ID*=1	Progressive increase of GBR traffic from t_1 to t_4	Average offered GBR load [Mb/s]	100	150	200	250	300	300	300	300	300
	Progressive increase of the number of non-GBR connections from t_5 to t_7	Average non-GBR load [number of DRBs]	10	10	10	10	10	20	30	40	40
Traffic demand variation on *RAN_Slice_ID*=4	Periodic reporting decrease due to smart meter application upgrades implemented in two phases (t_7 and t_8). No change in the number of connected devices.	Message periodicity [s]	120	120	120	120	120	120	120	60	30
Performance indicators	*RAN_Slice_ID*=1 – Admission Acceptance Ratio for GBR DRBs [%]		99.99	99.23	93.81	84.06	74.30	65.42	65.45	65.32	65.40
	RAN_Slice_ID=2 – Admission Acceptance Ratio for GBR DRBs [%]		99.97	99.99	99.93	99.97	99.99	99.94	99.92	99.96	99.91
	RAN_Slice_ID=1 – Average rate per Non-GBR DRB [Mb/s]		21.7	18.4	15.6	14.1	13.2	6.9	4.6	3.4	3.4
	RAN_Slice_ID=2 – Average rate per Non-GBR DRB [Mb/s]		21.7	18.4	15.6	14.1	13.2	11.9	11.9	11.8	11.9
	RAN_Slice_ID=3 – Average Message Queueing Delay [ms]		12.8	12.7	12.7	12.6	12.9	12.7	12.7	23.2	125.1
	RAN_Slice_ID=4 – Average Message Queueing Delay [ms]		52.4	52.3	52.4	52.4	52.2	52.3	52.4	62.5	165.5

Table 2. Illustrative assessment of the degree of isolation between the RAN slices.

A common frame duration of 10 ms is considered for the three L1 slices, while the slot duration within each slice depends on the used OFDM numerology. Note that with NR it is expected that a UE may be instructed to receive or transmit using a subset of the resource grid only [13]. Moving upward in the protocol stack, the two RAN slices (*RAN_Slice_ID*=1 and 2) on top of the *L1_Slice_ID*=1 are configured to use the same MAC instance (i.e. MAC-eMBB) and the same set of common logical channels (i.e. BCCH, PCCH and CCCH). Traffic isolation between *RAN_Slice_ID*=1 and 2 is enforced at both L2 and L3 levels through separate L2 and L3 slice descriptors. As with the two RAN slices for eMBB, the two slices for IoT services (*RAN_Slice_ID*=3 and *RAN_Slice_ID*=4) are also configured to use a shared MAC instance (i.e MAC-NB-IoT) and common logical channels. However, isolation is now only configured at the L3 level (i.e. both slices use the same L2 configuration descriptor). Finally, it can be noted that the Car IoT L1 slice is configured to be exploited only by *RAN_Slice_ID*=5, so that there is no need for isolation at L2 and L3.

Illustrative evaluation results are provided next to show how some L1, L2 and L3 configuration descriptor parameters impact the achieved performance and level of isolation across the RAN slices. The evaluation scenario considers an outdoor urban micro cell operating in the 3.4–3.8 GHz frequency band with a carrier bandwidth of 100 MHz (equivalent to 500 physical resource blocks [PRBs] for an OFDM numerology with Δf = 15kHz). Table 1 describes the simulation settings including traffic demand and configuration parameters of each RAN slice. Of note is that, regarding the two eMBB slices, the *Slice Authorized Capacity* parameter within L3 slice descriptors is used to limit both the maximum aggregated bit rate of the active GBR bearers allowed per slice and the average number of PRBs consumed by this traffic. Moreover, the parameter *Slice Resource Utilization* within L2 slice descriptors is used to limit the maximum share per RAN slice of the PRBs used to serve non-GBR bearers (i.e. *L1_Slice_ID*=1 resources not consumed by the GBR bearers). On the other hand, regarding the two mMTC slices, the parameter *Slice Authorized Capacity* is set to only limit the number of connected devices. On this basis, key performance indicators achieved for the eMBB and mMTC slices are given in Table 2, where t_0 indicates a time instant under the traffic demand conditions in Table 1 and subsequent time instants consider some traffic demand variations. As seen in Table 2, GBR traffic overload in *RAN_Slice_ID*=1 from t_1 onward is mitigated via admission control as soon as GBR traffic starts compromising the *Slice Authorized Capacity* configuration so that GBR traffic served through *RAN_Slice_ID*=2 remains unaffected. Note, however, that this GBR traffic increase leads to a reduction of the capacity left for non-GBR traffic and, consequently, non-GBR DRB rates are lowered in both *RAN_Slice_ID*=1 and *RAN_Slice_ID*=2. On the other hand, the increase in non-GBR connections in *RAN_Slice_ID*=1 from t_5 onward, while it reduces non-GBR DRB bit rates in *RAN_Slice_ID*=1, it does not hinder the non-GBR DRB performance in *RAN_Slice_ID*=2 thanks to the protection achieved through the L2 descriptor. Note also that IoT application performance remains unaffected from t_1 to t_6 because of the isolation enforced between L1 slices. Finally, it can be seen that the smart meter application upgrade in *RAN_Slice_ID*=4 leads to a delay increase for all IoT applications, including those

The proposed configuration descriptors provide a functional characterization of the RAN slice by parametrizing the features, policies and resources in place across L3, L2 and L1 protocol layers of the radio interface.

running over *RAN_Slice_ID*=3 since the only isolation mechanism between *RAN_Slice_ID*=3 and *RAN_Slice_ID*=4 is based on the limit of the number of connected devices, which remains unchanged. This situation could be avoided by either introducing L2 descriptors to limit the traffic served per slice or by dynamically upscaling *L1_Slice_ID*=2 from 1 PRB to 2 PRB and downscaling *L1_Slice_ID*=1 from 450 PRB to 449 PRB with a negligible impact on the eMBB slices.

Concluding Remarks

This article has described a framework that establishes the new blocks of information, configuration descriptors and extended protocol features to be introduced within a NG-RAN node for the realization of RAN slicing, thus enabling multiple RAN slices, with potentially different radio protocol behaviors as well as different levels of guaranteed network resources and isolation, to be concurrently multiplexed over the same cell. Remarkably, the proposed configuration descriptors provide a functional characterization of the RAN slice by parametrizing the features, policies and resources in place across L3, L2 and L1 protocol layers of the radio interface. An illustrative use case has been developed in detail to show the applicability of the proposed framework, visualizing the resulting radio interface protocol architecture along with the specific parameters involved in the RAN slicing configuration descriptors.

Acknowledgments

This work has been supported by the EU funded H2020 5G-PPP project 5G ESSENCE under the grant agreement 761592, and by the Spanish Research Council and FEDER funds under RAMSES grant (ref. TEC2013-41698-R).

References

[1] NGMN Alliance, "5G White Paper," Feb. 2015.
[2] 3GPP TR 22.864: "Feasibility Study on New Services and Markets Technology Enablers — Network Operation; Stage 1 (Release 15)," Sept. 2016.
[3] 3GPP TS 22.261 v15.0.0, "Service requirements for the 5G system; Stage 1 (Release 15)," Mar. 2017.
[4] 3GPP TS 23.501 V1.0.0, "System Architecture for the 5G System; Stage 2 (Release 15)," June 2017.
[5] 3GPP TR 28.801 V1.2.0, "Study on management and orchestration of network slicing for next generation network (Release 15)," May 2017.
[6] P. Rost *et al.*, "Mobile network architecture evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016, pp. 84–91.
[7] METIS II White Paper "Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design," Mar. 2016.
[8] O. Sallent *et al.*, "On Radio Access Network Slicing from a Radio Resource Management Perspective" *IEEE Wireless Commun.*, vol. 24, no. 5, Oct. 2017, pp. 166—74.
[9] X. Costa-Perez *et al.*, "Radio Access Network Virtualization for Future Mobile Carrier Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 27–35.
[10] 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, July 2016, pp. 32–39.
[11] V. Del Piccolo *et al.*, "A Survey of Network Isolation Solutions for Multi-Tenant Data Centers," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 4, Fourth Quarter 2016, pp. 2787–2821.
[12] 3GPP TS 38.300 V0.4.1, "NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)," June 2017.
[13] 3GPP TS 38.211 V0.1.0, "NR; Physical channels and modulation (Release 15)," June 2017.
[14] A. A. Zaidi *et al.*, "Waveform and Numerology to Support 5G Services and Requirements," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016, pp. 90–98.
[15] K. I. Pedersen *et al.*, "A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases," *IEEE Commun. Mag.*, vol. 54, no. 3, Mar. 2016, pp. 53–59.

Biographies

Ramon Ferrús (ramon.ferrus@upc.edu) is a tenured associate professor at UPC in Barcelona. His research interests include system design, functional architectures, protocols, resource optimization and network and service management in wireless communications. He has participated in multiple European Commission funded research projects as well as national research and technology transfer projects for public and private companies. He is the co-author of two books and 100+ papers published in peer-reviewed journals, magazines, conference proceedings and workshops.

Oriol Sallent is a professor at the Universitat Politècnica de Catalunya (UPC). He has participated in a wide range of European projects with diverse responsibilities as workpackage leader and coordinator partner, and he has contributed to standardization bodies such as 3GPP, IEEE and ETSI. He has published 200+ papers, 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 [S'98, M'04] is an associate professor in the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC) in Barcelona, Spain. He has been working in the field of wireless communication systems, with a particular focus on radio resource management cognitive radio networks and network optimization. He has been involved in different European projects and in projects for private companies. He has published more than 200 papers in international journals and conferences.

Ramón Agustí is a full professor at UPC in Barcelona. Since graduation he has been working in the field of digital communications on transmission and development aspects in digital radio. For the last 20 years he has been mainly concerned with aspects related to radio resource management in mobile communications. He has published more than 200 papers in these areas and has co-authored three books. He has been a member of the Spanish Engineering Academy since 2009.