Architecting RAN Slicing for URLLC: Design Decisions and Open Issues

Sergio Martiradonna*§, Andrea Abrardo†§, Marco Moretti‡§,Giuseppe Piro*§, and Gennaro Boggia*§
*Dept. of Electrical and Information Engineering - Politecnico di Bari, Bari, Italy
Email: {name.surname}@poliba.it
†Dept. of Information Engineering and Mathematical Sciences - Università degli Studi di Siena, Siena, Italy
Email: abrardo@diism.unisi.it
‡Dept. of Information Engineering - Università degli Studi di Pisa, Pisa, Italy
Email: marco.moretti@unipi.it
§CNIT, Consorzio Nazionale Interuniversitario per le Telecomunicazioni

***Abstract*—The Fifth Generation of mobile networks is emerging as a key enabler for Ultra Reliable and Low Latency Communications. However, to effectively design and provide safety-critical applications through mobile systems, many research issues still need to be deeply investigated. The most important ones include: (1) the dynamic and flexible management of radio resources of a new Radio Access Network jointly used by many virtual mobile operators, (2) the optimized and real-time configuration of network slices, and (3) the harmonious integration of Multi-access Edge Computing services. Starting from the efficient methodologies and solutions available in the current state of the art, this position paper sheds some important basis for the design of a comprehensive architecture enabling Radio Access Network slicing for Ultra Reliable and Low Latency Communications, including design criteria, system components and their baseline interactions, and critical open issues to be investigated in future research initiatives.**

***Index Terms*—URRLC, RAN slicing, MEC, architecture design, open issues**

I. Introduction

In the context of Fifth Generation (5G) mobile networks , Ultra Reliable and Low Latency Communication (URLLC) is providing the great opportunity to effectively implement safety-critical systems. Nevertheless, these applications rise groundbreaking challenges in terms of latencies, reliability, availability, and security [1], hence requiring a flexible Radio Resource Management (RRM) approach. However, it emerges as a very difficult problem to solve. Luckily, the emerging network slicing paradigm poses a strong basis to address this fundamental research challenge [2].

The most common slicing scenario includes a single Infrastructure Provider (IP) leasing its network resources to a set of independent Mobile Network Operators (MNOs) that provides advanced network services [3]. For each slice, the IP identifies the number of resources that can be used by the MNOs to provide different services to their subscribers. Moreover, MNOs should have the ability to adapt their slice requests to their users' requirements in real-time, dodging additional expenses due to the problem of resources overbuying. Thus, the slice request generation, i.e., when each MNO declares its desired slice configuration to the IP, clearly becomes crucial [4]. The inherent requirements of users demanding URLLC service, as well as their high mobility, put further constraints.

Starting from the valuable methodologies and solutions available in the current state of the art (see Section II and III for more details), this position paper sheds some important basis for the design of a comprehensive architecture enabling Radio Access Network slicing for Ultra Reliable and Low Latency Communications. Specifically, it presents a high-level architecture jointly tackling the problems related to URLLC and network slicing by means of a Multi-access Edge Computing (MEC) system, while posing particular attention to design criteria, system components, and their baseline interactions. The conceived architecture grounds its roots on the Platform as a Service (PaaS) paradigm: it is assumed that the IP is the only entity allowed to manage radio resources and network slices. To provide a further insight, some critical open issues are discussed too, which inevitably pave the way towards future research initiatives in this direction.

The remainder of this paper is as follows. Section II discusses the background on network slicing. Section III reviews methodologies and solutions for the radio resource management for URLLC. The proposed RAN slicing architecture is presented in Section IV. Finally, Section V draws the conclusions.

II. Background on Radio Access Network Slicing

The concept of network slicing is inherently related to the possibility of performing service differentiation. At the Radio Access Network (RAN) layer this concept can be captured by the mapping of Quality of Service Class Identifiers (QCI) to data radio bearers, hence allowing the creation of end-to-end chains where different traffics are treated according to different policies, at the same time satisfying a wide range of heterogeneous requirements. However, satisfying all these heterogeneous constraints operating at a per-flow level is simply not possible, due to some fundamental technical limits [3]. The vision of network slicing is expected to open new business models for IPs. For instance, a mobile operator will

978-1-7281-2923-5/19/$31.00 ©2019 IEEE

be able to split its physical network resources into multiple logical slices and lease these slices to interested parties. Besides, IPs may expand their business to the continuously growing market of flexible, high performance and on-demand network deployment for differentiated service typical of 5G networks.

The static partitioning of the resources to create different slices, despite being straightforward, leads to low efficiency [3]. Moreover, it is not adequate to enforce the full concept of network slicing, where different operators wish to have - at least partial - control of the underlying infrastructure, which they can configure and operate independently. Following this line of reasoning, in [5] the concept of the 5G Network Slice Broker is introduced, which enables new players to request and lease resources from IPs dynamically via well-defined interfaces. An interesting holistic vision of RAN slicing addressing the above-mentioned considerations, is presented in [4]. In [6] an overall analysis of the RAN slicing problem in a multi-cell network is presented and four different slicing approaches working at different RRM functionalities levels are proposed. The problem of slice enforcement expressed as Resource Blocks (RBs) allocation to each slice for a given resource partitioning policy among the admitted slices in presented in [7], addressing, in particular, the isolation problem. A distributed approach for performing RAN slice formation among competing MNOs is presented in [8] and [9]. Some other papers related to RAN slicing, e.g., see [5], [10], [11], and [12] envisage an architecture where the control part of the RAN is partially in charge of a third party entity, generally referred to as tenant, that is responsible for making slice formation requests and for operating slice enforcement using open interfaces to communicate with the IP. Finally, the problem of providing URLLC using 5G network slices is addressed in [1], [13]–[15].

III. Radio Resource Management for URLLC

RRM for URLLC should *i)* minimize latency, *ii)* maximize reliability by favoring robust modulations over high data rate modulations, and *iii)* optimally allocate power and time-frequency resources among different users. Effective RRM is therefore a complex task for a number of reasons.

- Diversity, which is critical for achieving high reliability, can only be gained by spreading the transmissions over multiple frequency channels.
- The existence of heterogeneous services, multiple Quality of Service (QoS) requirements, and diverse set of available resources, greatly reduce the tractability of the problem.
- The number of resources to be managed by the schedulers is a function of several parameters and the control of this huge amount of data within a given deadline requires very high computational resources.

Recently, there have been important advances in RRM for URLLC. Among several others, the studies in [16]–[19] jointly investigate enhanced Mobile BroadBand (eMBB) and URLLC schedulers. In parallel, the rapid advances in the Machine Learning field grant coping with the requirements imposed by ever more sophisticated RRM functionalities. At the same time, this kind of frameworks is gaining a more central role since it is capable of carrying out the prompt decisions required in 5G [20]–[22].

In the URLLC RAN slicing context, MEC can guarantee extremely low latency and bandwidth efficiency, differently from traditional centralized cloud computing, thanks to the use of some servers installed at the edge of the network (i.e., proximal to the end users) [23]. In essence, a virtualized application platform is built over physical hardware resources provided by the host machine that mounts the MEC server within the Cloud - Radio Access Network (C-RAN). This application platform offers middleware services to the applications running on the MEC server through Application Programming Interfaces (APIs) for regulating the communication between application and service and between the various applications, collecting and providing information about users and cell, and managing the routing of traffic to and from applications.

For these reasons, MEC servers deployment in the C-RAN appears as one of the best solutions for RAN slicing, guaranteeing both a strong versatility to extremely time-variant conditions and easiness of interactions between all the parties involved in the slicing operations.

IV. The Proposed Architecture

Although network slicing is expected to open new business models for all the interested parties, it is of the utmost importance to emphasize that roles must not change. As a consequence, IPs and third-party entities should not be aware of MNOs' most valuable information, as well as MNOs are not authorized to precisely comprehend the procedures and algorithms implemented by the IPs and the internal functioning of the provided apparatuses.

Based on these premises, we envision an architecture to favor collaboration while keeping interactions among all the parties involved clearly distinct. As a consequence, we assume that IP is the only entity allowed to implement resource allocation. In addition, for simplicity reason, we will assume the network of a single IP in the subsequent analysis. However, it is important to note that the proposed architecture can be easily scaled for multiple IPs.

A. Overview

We consider a general network scenario constituted by a New Generation Service Oriented Core and a C-RAN. In particular, C-RAN architecture is used on the RAN side in order to implement real time functions, RAN slicing, on-demand deployment of resources, and flexible coordination. Moreover, we choose to equip the C-RAN with the Mobile Cloud Entity (MCE) to implement functions requiring high real-time performance and computing load based on different service requirements and resource configurations. According to the IP implementation, C-RAN can have different levels of control of radio functions, from an overall control of the gNBs to the complete Radio Remote Heads capabilities. RAN

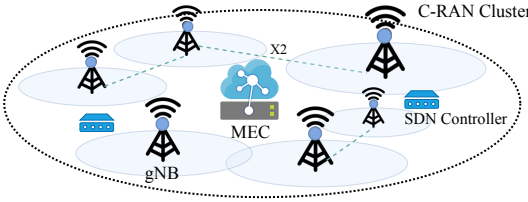

Fig. 1. C-RAN Cluster with multiple gNBs served by a single MEC server and two SDN Controllers.

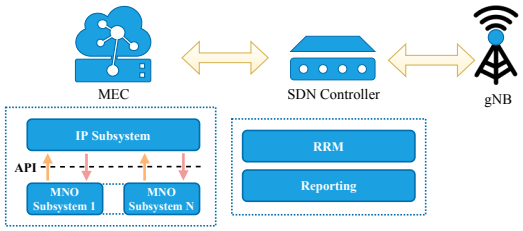

Fig. 2. Functions and interactions between the elements of the proposed architecture.

is composed of numerous gNBs settled in different geographic areas. Nearby gNBs are then grouped in a C-RAN cluster, a spatially isolated and self-sufficient logical network partition. One or several MEC servers are then bootstrapped to each C-RAN cluster and connected to one or more Software Defined Networking (SDN) controllers supporting the underlying physical infrastructure (see Figure 1).

In order to automatically generate network slicing services according to the specified requirements, two different entities, namely IP Subsystem and MNO Subsystem, are envisaged to interact with each other. In this scenario, the IP dynamically leases computing resources for granting MNO Subsystems to be virtualized on MEC, according to the *PaaS* paradigm. A high-level overview of this architecture is depicted in Figure 2. The main reason why the proposed architecture is built on top of the *PaaS* paradigm lies in the poorer performance of a *Infrastructure as a Service (IaaS)*-based architecture, in which the IP only leases its resources and the MNOs are left in charge of developing their own virtualized subsystems. However, we remark that the envisioned network scenario is flexibly suitable to both *IaaS* and *PaaS* architecture deployments.

B. IP Subsystem

It creates different RAN slices according to the control directives coming from the MNO Subsystems, which are first translated in order to be effectively managed. Here, the IP provides specific APIs for the submission of slice requests. The complex directives are then turned into simpler records containing all the important parameters — e.g., QoS requirements, frame structure, transmission configuration, etc.. The resource requirements are checked in order to determine whether a slice should be admitted or not. Admission control is thus performed on a per-slice basis by taking into account the Service Level Agreements (SLAs) with the MNOs as well as the reporting information from the gNBs. To this end, the IP Subsystem communicates with the SDN controllers managed by the C-RAN sending network-level RAN performance requirements necessary to create and/or enforce already present RAN slices. It is of critical importance, especially for URLLC services, to determine which parameters the MNOs and IP are allowed to share, as well as the level of control the MNOs can have on the C-RAN. In fact, the variability of the service requirements may call for feedback messages between the IP and the MNOs Subsystems (or toward third-parties) leading to an immediate slice renegotiation. After having formed a URLLC slice, it may happen that suddenly the channel conditions have changed so much, or that the MEC server is overloaded, that no enforced RRM algorithm can guarantee the satisfaction of the constraints provided by the resource allocation policy established for that slice at the time of its creation.

C. MNO Subsystem

MNO Subsystems mainly formulate slice requests by specifying both general information (e.g., the time duration of the slice, type of services to be provided) and high-level control directives in order to successfully address the requirements for the requested slice. Requests are forwarded to the IP Subsystem through the provided APIs. In addition, each MNO Subsystem has to continuously monitor customer requirements based on current network status in order to check whether SLA violations happen and to properly reconfigure the parameters included in the slice requests. Based on the previous behavior of the IP Subsystem, if multiple slice requests need to be forwarded, the MNO Subsystem is in charge of the assignment of inter-slice priorities according to the required SLAs. For this purpose, it is necessary to be informed in advance and to have the faculty both to renegotiate the slice and to decide to interrupt other flows, if they have less priority.

Handling of RAN slice requests submitted by third-party entities should be supported as well. Vertical industries subscribe a specific plan with MNOs who are then completely in charge of managing the sliced networks according to the agreed decisions.

D. Open Issues and Future Research Activities

The identification of the readiness level of the underlying technologies paves the way towards the future research on the reference themes. The most relevant issues that affect RAN slicing for URLLC as well as interesting research activities to address in the future are listed in what follows.

- Appropriate APIs between IP and MNO Subsystems are a key feature for boosting the performance and achieving extreme flexibility.
- MEC Servers utterly need to be protected from security threats.
- It is of the utmost importance to properly dimension the performance reporting between the IP and MNOs (e.g., periodical, triggered by events, or both).

- The architecture should consider different time scales for slice generation since 5G is a heterogeneous context.
- It is imperative to design efficient slicing enforcement algorithms to prevent interference between different MNOs, i.e., for providing inter-slice isolation.
- IP Subsystem should be able to manage advanced multi-cell techniques, e.g., Coordinated Multi-Point, beamforming, etc..
- Granularity constraints in spectrum and radio-level resource sharing are a primary concern.
- The high mobility of users put further constraints both on the RRM and the slicing problem.
- The combination of different technologies (e.g., Non Orthogonal Multiple Access (NOMA), shorter Transmission Time Intervals (TTIs), mini-slots, distributed machine learning framework, ultra-lean design, grant-free transmissions, flexible TDD configurations, device-to-device communication, delay-budget reporting, etc..) is beneficial for URLLC services [24].
- Net Neutrality should always be guaranteed.

V. Conclusion

In this paper, valuable methodologies and solutions currently available in the literature for carrying out effective RRM procedures in the context of URLLC for 5G networks are presented and discussed. Starting from this analysis, we then give the basis for the design of a comprehensive architecture enabling Radio Access Network slicing for Ultra Reliable and Low Latency Communications. Specifically, we present a high-level architecture jointly tackling the problems related to URLLC and network slicing by means of a MEC system, while posing particular attention to design criteria, system components, and their baseline interactions. The conceived architecture grounds its roots on the PaaS paradigm: it is assumed that the Infrastructure Provider is the only entity allowed to manage radio resources and network slices. To provide a further insight, some critical open issues are discussed too, which inevitably pave the way towards future research initiatives in this direction.

Acknowledgment

This work was supported by the project "Pre-commercial trials of 5G technology using spectrum in the 3.6 GHz-3.8 GHz range - Area Milano", founded by the Italian MISE, by the PRIN project no. 2017NS9FEY entitled "Realtime Control of 5G Wireless Networks: Taming the Complexity of Future Transmission and Computation Challenges", and PON project INTENTO (36A49H6), both founded by the Italian MIUR.

References

[1] L. Zanzi and V. Sciancalepore, "On Guaranteeing End-to-End Network Slice Latency Constraints in 5G Networks," in *Proc. IEEE International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018.

[2] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, May 2017.

[3] Ö. U. Akgül, I. Malanchini, and A. Capone, "Dynamic Resource Trading in Sliced Mobile Networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 220–233, March 2019.

[4] S. D'Oro, F. Restuccia, and T. Melodia, "Toward Operator-to-Waveform 5G Radio Access Network Slicing," *arXiv:1905.08130 [cs]*, 2019.

[5] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From Network Sharing to Multi-Tenancy: The 5G Network Slice Broker," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 32–39, 2016.

[6] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, 2017.

[7] S. D'Oro, F. Restuccia, A. Talamonti, and T. Melodia, "The Slice Is Served: Enforcing Radio Access Network Slicing in Virtualized 5G Systems," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, April 2019, pp. 442–450.

[8] S. D'Oro, F. Restuccia, T. Melodia, and S. Palazzo, "Low-Complexity Distributed Radio Access Network Slicing: Algorithms and Experimental Results," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2815–2828, Dec 2018.

[9] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, "Network Slicing Games: Enabling Customization in Multi-Tenant Mobile Networks," *IEEE/ACM Transactions on Networking*, vol. 27, no. 2, pp. 662–675, April 2019.

[10] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, "Dynamic Network Slicing for 5G IoT and eMBB services: A New Design with Prototype and Implementation Results," in *Proc. IEEE Cloudification of the Internet of Things (CIoT)*, July 2018.

[11] Y. Sun, M. Peng, S. Mao, and S. Yan, "Hierarchical Radio Resource Allocation for Network Slicing in Fog Radio Access Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3866–3881, April 2019.

[12] K. Zhu and E. Hossain, "Virtualization of 5G Cellular Networks as a Hierarchical Combinatorial Auction," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2640–2654, Oct 2016.

[13] A. Ksentini, P. A. Frangoudis, A. PC, and N. Nikaein, "Providing Low Latency Guarantees for Slicing-Ready 5G Systems via Two-Level MAC Scheduling," *IEEE Network*, vol. 32, no. 6, pp. 116–123, November 2018.

[14] T. Guo and A. Suárez, "Enabling 5G RAN Slicing With EDF Slice Scheduling," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2865–2877, March 2019.

[15] D. Tang, C. Hu, and T. Dang, "Delay-Aware Resource Allocation for Network Slicing in Fog Radio Access Networks," in *Proc. IEEE International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct 2018.

[16] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[17] A. Anand, G. De Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, April 2018, pp. 1970–1978.

[18] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, "eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach," *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, April 2019.

[19] J. Park and M. Bennis, "URLLC-eMBB Slicing to Support VR Multimodal Perceptions over Wireless Cellular Systems," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2018.

[20] G. Zhu, J. Zan, Y. Yang, and X. Qi, "A Supervised Learning Based QoS Assurance Architecture for 5G Networks," *IEEE Access*, vol. 7, pp. 43 598–43 606, 2019.

[21] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, and P. Soldati, "Learning Radio Resource Management in 5G Networks: Framework, Opportunities and Challenges," *arXiv:1611.10253 [cs]*, 2017.

[22] A. Azari, M. Ozger, and C. Cavdar, "Risk-Aware Resource Allocation for URLLC: Challenges and Strategies with Machine Learning," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 42–48, March 2019.

[23] ETSI, "MEC in 5G Networks," White Paper No. 28, June 2018.

[24] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.