Impact of Slice Function Placement on the Performance of URLLC with Redundant Coverage

Abdellatif Chagdali, Salah Eddine Elayoubi, Antonia Maria Masucci

Impact of Slice Function Placement on the Performance of URLLC with Redundant Coverage

Abdellatif Chagdali*‡, Salah Eddine Elayoubi‡, Antonia Maria Masucci*
* Orange Labs, Châtillon, France
‡ Université Paris Saclay, CentraleSupélec, CNRS, Gif-Sur-Yvette, France

***Abstract*—Network slicing has emerged as a promising technical solution to ensure the coexistence of the various 5G services. While the evolution of the 5G architecture for supporting slicing has been thoroughly studied, the impact of the architectural options on RAN resource allocation efficiency is still unclear. This article fills a gap in this area by assessing the impact of architecture choices on the quality of service of different services, with a focus on ultra-reliable low-latency communication applications. We propose architectural options based on the placement of the entities responsible for implementing these functions. We then assess their impact on the radio resource allocation flexibility when slices span two radio access technologies with redundant coverage. Our numerical experiments show that the placement of the slice management functions plays a crucial role in the choice of the radio resource allocation scheme that best fits URLLC slices.**

***Index Terms*—5G, network slicing, quality of service, ultra-reliable low-latency communication (URLLC), resource allocation, redundancy, join the shortest queue.**

I. Introduction

The fifth generation of mobile networks (5G) differs from its predecessors as it harbors a novel and unprecedented service-oriented vision along with the evolutionary view. As a matter of fact, 5G systems will not provide just an increase in data rates but will also define new use cases in hand with legacy services. Delivering these new services requires a versatile, scalable, efficient, and cost-effective network, capable of accommodating its resource allocation to act upon the heterogeneous nature of demands. For instance, Enhanced Mobile Broadband (eMBB) services necessitate high data rates, wide-area coverage, and high-user density. On the other hand, Ultra-reliable and Low-Latency Communications (URLLC) define strict requirements in terms of latency and packet loss.

Network slicing has emerged as one of the fundamental concepts proposed to raise the efficiency and provide the required plasticity of 5G mobile networks. The idea is to provision resources for different vertical industries by building multiple End-to-End (E2E) logical networks over a shared infrastructure. Each network slice (*i.e.* logical network) is customized to deliver a specific service to a tenant.

Even though the concept of network slicing is relatively new, the corresponding literature that deals with it is already rich, especially on architecture and management aspects. For instance, [1] offers a holistic approach by discussing the management and orchestration for E2E slices, including infrastructure layer, network function layer, and service layer. In paper [2], the authors discuss the architectural concepts for slicing, including the mapping of network functions (NFs) to satisfy the discordant performance targets of 5G use cases. While network slicing is an E2E concept, most of the research has focused on core slicing leading to mature architecture proposition powered by the emergence of cloud computing, Network Function Virtualization (NFV) and Software-Defined Networks (SDN) [1], [3], [4].

However, Radio Access Network (RAN) slicing introduces a distinct set of issues compared to core slicing. Third Generation Partnership Project (3GPP) foresees novel Radio Access Technologies (RATs), new subcarrier spacing, and frame structure to provide RAN adaptability, given the dissonant nature of verticals' demands [5]. Thus, a key challenge is to choose the propitious RATs for each service during the preparation phase of RAN slice creation. Besides, it's indispensable to design isolation mechanisms between RAN slices along with a charging framework that takes into consideration the role of third-party players.

The authors in [6] focus on slicing implementation in the RAN and identify its enablers, among which we can find: flexible numerology, mobile edge computing, and slice tiling. The latter arranges time-frequency resources with the same numerology in resource block groups that the scheduler allocates to the adequate slice type; for optimal resource allocation. In [7], the authors probe into RAN slicing management in a multi-cell network by studying four architecture proposals for radio resource sharing and discuss their different levels of granularity, isolation, and customization aspects. Moreover, the authors in [8] advocate for radio protocol layer descriptors that outline the features, policies, and resources needed to create and customize multiple RAN slices.

While these works paved the way for the definition of slicing concept in 5G, they did not tackle the impact of the 5G ecosystem openness to new actors on resource allocation implementation in the RAN. Indeed, even if the 5G New Radio (NR) has been designed as highly flexible to ensure efficient multiplexing between slices, the task of radio and computing resource allocation is still cumbersome. The multiplication of actors that have stakes in the RAN makes it arduous to allocate resources to the slices, as the resources supposedly belong to multiple Infrastructure Providers (InPs). The latter InPs contract Service Level Agreements (SLAs) with different Mobile Service Providers (MSP) and verticals. Our objective in this paper is to build on the concepts developed in the literature, to construct a Radio Resource Management (RRM) framework for URLLC and eMBB slices. We specifically:

Function	Location	Functionality	Owner	Autonomy
UE scheduler	UE	Dispatches UE traffic to access points	Vertical	Applies policies specified by the vertical
BS scheduler	Base station	Allocates time/frequency resources to UEs	InP	Applies policies specified by the InP
NSSMF	RAN (e.g. Cloud RAN)	Orchestrates RAN resource allocation to slices	InP	Defines policies for the InP base stations
NSMF	MSP management server	Defines traffic steering policies for the slice	MSP	Defines MSP policies
CSMF	Tenant management entity (e.g. application Server)	Updates slice requirements and SLAs	Tenant	Defines tenant policies and needs

TABLE I: Entities involved in RAN resource allocation and their roles.

1) identify the different actors that have stakes in RAN resource allocation and their ownership of the different slice and network functions,
2) propose a placement for the intelligent entities that take decisions on traffic steering and resource allocation, for all the involved actors,
3) quantify the impact of the defined architectural options on the Quality of Service (QoS) in the practical case of redundant coverage of two RATs.
4) compare the performances of different packet replication schemes, namely Join-the-Shortest-Queue and systematic redundancy, for different architectural options.

The remainder of this paper is organized as follows. Section II describes slice management functions and stipulates their ownership and roles in the radio resource management. Section III presents the impact of the architectural options for the placement of these functions on the possible scheduling mechanisms for URLLC in an industrial scenario. Section IV compares the performances of these scheduling mechanisms and shows the best policies in each of the studied scenarios. Section V eventually concludes the paper.

II. RAN RESOURCE ALLOCATION AND TRAFFIC STEERING

Before describing the slice management function placement options, we aim in this section at identifying the role of each player in the slice management and the entities that are responsible for managing traffic and resources for the slices.

A. Slice management functions description and ownership

The MSP has to create and simultaneously maintain many Network Slice Instances (NSI). An NSI is composed of Core Network (CN) and RAN Network Slice Subnet Instances (NSSI), arranged to provide necessary resources and functionalities and thus deliver the tenants' services. Each NSSI encompasses Physical Network Functions (PNFs) and Virtual Network Functions (VNFs) that are either dedicated or common among different slices. According to the solution advocated by 3GPP and illustrated in Figure 1, the tenant's management function called the Communication Service Management Function (CSMF) forwards the service requirements to the Network Slice Management Function (NSMF). Then, the NSMF translates the E2E high-level performance requirements desired by the tenant to CN and RAN low-level requirements managed by the Network Slice Subnet Management Function (NSSMF). Thereafter, the RAN NSSMF converts the low-level requirements into RRM specific requirements and sets the resource allocation policy at the MAC scheduler of the Base Station (BS), whereas the CN NSSMF deploys and maps the service-oriented VNFs. Both the RAN and CN NSSMFs send periodic performance reports to the NSMF so that it can verify that the service requirements are respected. For example, if the RAN NSSMF violates the latency requirement of a network slice, the NSMF can adjust the scheduling policy by reserving more resource blocks. It can also alter the admission control procedure by rejecting any other network slice requests as long as the served slices SLAs are not respected. Table I summarizes these entities, their owners, and their roles in the slice resource allocation.

Note that the resource allocation task is particularly complicated in case several MSPs lease resources from multiple InPs. Indeed, the NSMF belongs to the MSP and has as objective to ensure that the tenant's SLA is respected. Nevertheless, there is a RAN NSSMF that belongs to each InP, which has control over the resources of this particular InP only, as illustrated in Figure 1. In the latter, we consider three slices belonging to three different tenants. For each slice, we deploy an NSSMF per InP RAN. The question here is how to design and implement resource management policies in such a distributed architecture while taking into consideration tenant, MSP, and InP perspectives.

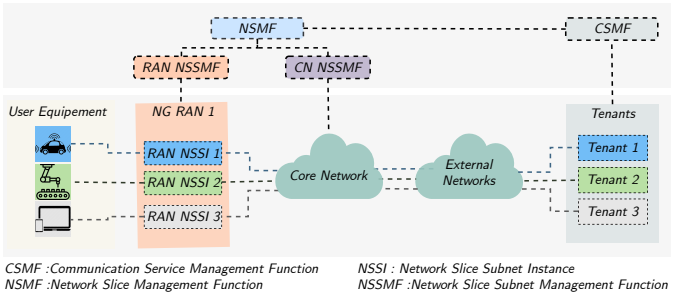

Fig. 1: Overview of the infrastructure and management Layer in a network slicing scenario.

B. Resource allocation from MSP perspective

From the MSP perspective, the NSMF translates the tenant requirements into a traffic steering policy that determines to which InP(s) the packets of a specific User's Equipment (UE) are to be forwarded. Such a policy may be generic, *e.g.* to privilege a particular InP when possible. Alternatively, it can be context-aware, which means examining the instantaneous load of the base station pertaining to an InP and its radio conditions with respect to the UE. For example, a potential policy is to connect a particular UE to a single InP, or to split its packets between several InPs, or even to duplicate them to increase reliability. Specifically, the NSMF can, for example,

decide that 70% of generated packets go through the main InP while the remaining 30% go through secondary InP during the validity time of the high-level policy.

In case the MSP applies the decided policy without coordination with the InPs, an entity hosted in the UE, capable of implementing the NSMF scheduling strategy, is required. Otherwise, the traffic steering policy can be implemented as a shared NF among multiple slices on the InPs infrastructure, or as dedicated NF with some cooperation between slices to meet the heterogeneous optimization targets and for effective use of the radio resources [8].

C. Resource allocation from tenant perspective

From the tenant perspective, the CSMF determines dynamically the amount of resources that need to be allocated to the slice for continuously respecting the SLA, knowing the current traffic demand. In order for these requests to be accurate, the CSMF has to rely on the information originating from the application server and/or from the end-users. The time scale for these traffic reports has to be larger than the actual scheduler time scale, *i.e.* in the order of tens of seconds. In the specific case where the tenant is a "big" vertical that can deploy its own infrastructure (e.g., railway and highway management companies), it has the ability to bypass the MSP and acquire the resources directly from InPs, having thus the same behavior of MSPs, described previously.

D. Resource allocation from InP perspective

From the InP perspective, the NSSMF receives the resource allocation requests from the UEs belonging to different slices and applies some scheduling/admission control policies to them. The devised policies of the InP have to dynamically share the resources among the slices to raise the overall resource efficiency, especially that leasing fixed shares of resources will limit the multiplexing gains.

Note that, from an InP perspective, [9] introduces the so-called 5G network slice broker, hosted in the NSSMF of the InP, that gathers global network load measurements and configures the RAN scheduler policies based on the negotiated SLA and the size of the network slice. Moreover, the openness of the mobile network may lead to an adversarial behavior of MSPs consisting of maximizing the acquired share of resources. In order to deal with this issue, a share-constrained proportional allocation mechanism is exploited in [10], and the share obtained by each tenant is determined by the equilibrium point of a network slicing game. In the same context, the authors in [11] investigate resource allocation mechanisms between tenants using game theory tools to model the non-cooperative behavior of slices. However, these works are limited to multiple tenants sharing a single InP infrastructure.

III. Impact of Placement of Intelligent Entities on Radio Resource Allocation for Slices

We now study the placement of entities in charge of resource allocation in light of the above description of the slice management functions. We consider, for illustration, the case of a smart factory where several base stations (5G NR and/or legacy) are deployed to establish a redundant coverage, essential for ensuring URLLC QoS[1], as illustrated in Figure 2. The tenant may own and manage some small cells deployed within the factory, while the InP manages base stations, operating in the sub 2 GHz spectrum for ensuring full coverage. While some UEs will be covered by the macro cells only, it is envisioned that most locations will be covered by at least two overlapping cells, providing flexibility in resource allocation and redundancy for ensuring reliability. We hereafter display three potential resource allocation schemes exploiting these advantages.

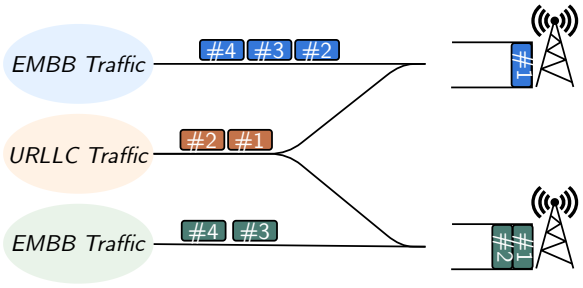

Fig. 2: Traffic steering in industry 4.0 use case. Two base stations in the neighborhood of URLLC user equipment.

A. Intelligence placed at the level of a shared RAN NSSMF

In this case, the traffics of all RAN slices and the radio resources of all base stations are managed via a shared RAN NSSMF with a single compound MAC scheduler. The latter has access to real-time information concerning the time-frequency matrix of each base station, thus allowing grant-based resource allocation. This case is enabled when the resources of all base stations are centrally managed within a common Cloud-RAN linked to the base stations by a high capacity fronthaul, as illustrated in Figure 3a.

A dynamic strategy can thus be applied by the NSSMF to URLLC traffic, which consists of sending packets to the BS with the lowest instantaneous load in order to minimize latency. As of eMBB traffic, it is served by one of the two base stations independent of the instantaneous load, *i.e.* each base station has its own eMBB traffic to serve and manages the URLLC traffic jointly with the other base station, as illustrated in Figure 2. This strategy can be applied in both uplink and downlink; it is straightforward in the downlink where the application server sends the URLLC packets to the NSSMF that directs them to the adequate base station for transmission. As of the uplink, the UE sends a scheduling request to NSSMF that issues a scheduling grant on one of the base stations. Consequently, the uplink case is more challenging as this control process may introduce latency between the moment the loads are observed by the NSSMF and the moment the scheduling grant is issued for the URLLC user.

[1]5G NR and 4G base stations can natively cooperate via a common core network, whereas [12] prescribes the Non-3GPP Interworking Function (N3IWF) for combining accesses using proprietary or WiFi technology.

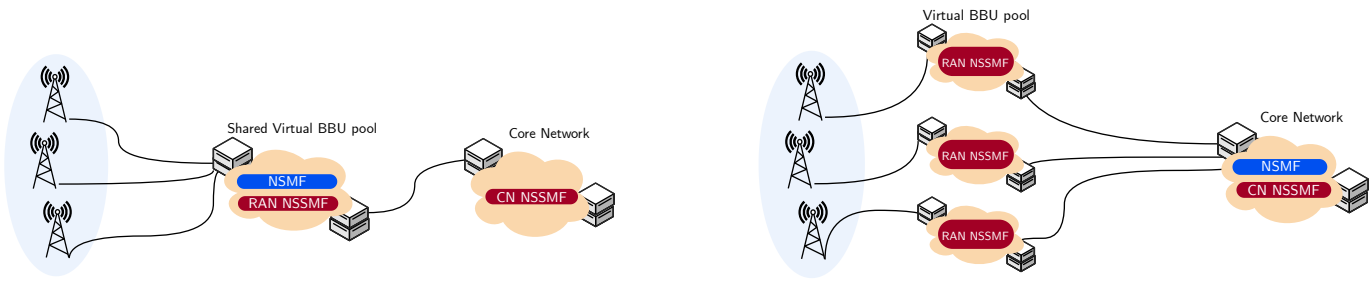

Fig. 3: Distribution of Management Functions in a factory scenario. (a) Intelligence placed at the level of a shared RAN NSSMF; (b) Intelligence placed at the NSMF level.

B. Intelligence placed at the NSMF level

When there is restricted coordination between the InPs, and between the MSP and the InPs, as in the case where each base station has its own Baseband Units (BBU), loosely linked to other BBUs, performing intelligent steering of each packet based on the instantaneous load of each cell is difficult to achieve. This is illustrated in Figure 3b. In this case, a long-term policy (e.g. based on a time granularity of tens of seconds) is to be applied, managed by the NSMF located somewhere at the level of the core network. For this policy to be effective, the UEs (in the uplink) and the application server (in the downlink) have to apply the policy provided by the NSMF on a packet basis, but without further information on the instantaneous load of each cell. When the decision about the destination of the packet is taken, the remainder of the scheduling process is performed in a classical way, and the RAN NSSMF does not need to know about the slice policy. We consider hereafter two feasible policies for URLLC:

1) Long-term traffic steering with no redundancy: It entails the division of the URLLC traffic proportionally, based on the base stations' average capacities as estimated by the NSMF, or as provided to the MSP by the RAN NSSMF of each InP.
2) Long-term traffic steering with redundancy: In the absence of any information about the capacities of the different base stations, and in order to ensure reliability, redundancy is a costly yet simple strategy. This implicates sending systematically the arriving URLLC packets to both base stations. While packet redundancy can achieve high reliability as it enables the experience of minimum queuing latency between the BSs, it leads to the under-utilization of radio resources. The NSMF broadcasts the policy to the URLLC user equipment during the slice instantiation.

IV. Performance Evaluation

First, we simulate the system presented in Figure 2, with only URLLC packets, via a resource reservation scenario for URLLC slice. Then, we consider a use case where eMBB and URLLC slices share the same resources. The goal is to obtain quantitative insights on the impact of architectural consideration on delivering the hard latency requirement of the URLLC service for different architecture options.

We consider three Poisson processes of arriving packets. URLLC packets are steered with regard to the network architecture and the placement of resource management entities (see Figure 3). The mean number of packets generated by URLLC users is 100 packets/s. When URLLC and eMBB slices share the same resources, we assume that eMBB users are sending packets with two independent processes for the two base stations, as illustrated in Figure 2. The eMBB packets' size is set to 1500 bytes, whereas the size of the URLLC packets is 32 bytes. Motivated by the flexibility of the 5G NR air interface, we consider that URLLC packets are served on a mini-slot basis, of 2 OFDM symbols, while eMBB packets are served on a legacy 1 ms TTI [13]. Service times of URLLC and eMBB packets depend on the used modulation and coding scheme. The latter varies from one user to another, depending on its average radio conditions. Hence, we set the uplink and downlink average spectral efficiency for eMBB packets equal to 6.75 and 9 bits/Hz/s, respectively [14]. Additionally, we model the uplink and downlink spectral efficiency for URLLC packets as a discrete uniform distribution over the set $\{1, 1.5, 2, 2.5\}$ bits/Hz/s.

We evaluate through Monte Carlo simulations the outage probability of URLLC traffic originating from the placement of the intelligent entities, for both cases discussed above. We send packets for a time window of 50 seconds, and we carry the same experiment 100 times. The outage probability is defined as the probability that the packets' latency exceeds a predefined threshold, which depends on the target latency. We set the target latency threshold for URLLC to 0.5 ms.

Simulation parameters	**Value**
URLLC packet size	32 bytes
eMBB packet size	1500 bytes
Control plane reports	100 μs
Latency threshold	0.5 ms
Total bandwidth for BS1 (resp. BS2)	10 Mhz (resp. 20 MHz)
URLLC reservation for BS1 (resp. BS2)	1 MHz (resp. 2 MHz)
URLLC packet generation per user	100 packets/s
Spectral efficiency for URLLC	$\{1, 1.5, 2, 2.5\}$ bits/Hz/s
UL (DL) average spectral efficiency for eMBB	6.75 (resp. 9) bits/Hz/s

TABLE II: Parameters for performance evaluation

A. Bandwidth Reservation Case for URLLC Slice

First, we study the impact of slicing architecture on URLLC traffic. In order to isolate them from eMBB traffic, we reserve

for this type of traffic a sub-band on each base station. In this setting, we assume a reserved bandwidth of 1 MHz and 2 MHz on BS1 and BS2, respectively. In Figure 4, we plot the URLLC traffic outage probability stemming from the different policies while varying the number of URLLC users. We consider two procedures and their variants:

- Decision in a shared RAN NSSMF: When the scheduling decision is taken at the RAN NSSMF level, the packet steering policy depends on the load of the base station. We consider two practical variants. The first variant supposes the knowledge of the instantaneous load with no delay. The second case takes into account the control plane signaling delay, especially for the uplink, where the user terminal relies on information sent by the NSSMF some time ago (100 μs in the numerical example) to make its decision.
- Decision in a far NSMF: When the instantaneous load is not available as the decision is taken at the NSMF level, we consider two variants. First, an NSMF proportional traffic steering based on the average capacities ($1/3$ of the traffic over BS1 and $2/3$ over BS2 in our case) is implemented; this corresponds to probabilistic routing of the packet based on long-term policies sent by the NSMF. Second, we consider the instance where each packet is duplicated to the two interfaces for profiting all the time from the interface with the lowest load.

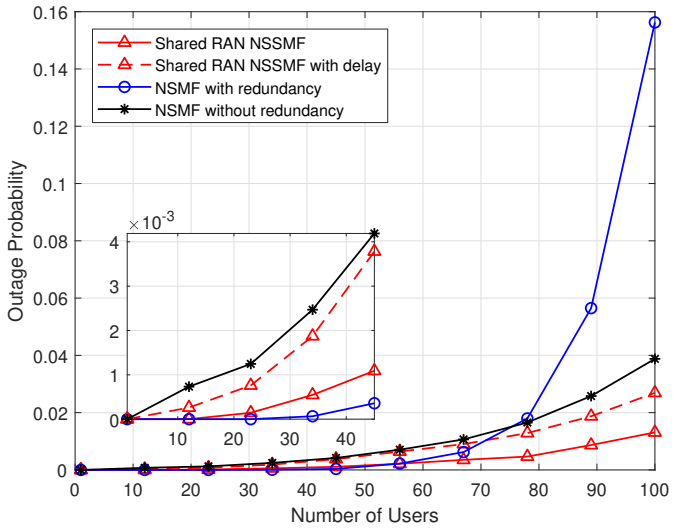

Fig. 4: Outage probability in the case of bandwidth reservation for URLLC packets.

When having a global look at the performance, we observe two regimes, each giving an advantage for one of the architectural options. First, in medium to high load regimes, placing the intelligence at a shared RAN NSSMF has a clear advantage over the management in a far NSMF entity. While these regimes are classically targeted for eMBB, they are not suitable for URLLC where a very low outage probability is sought, in the order of 10^{-6} to 10^{-5}, corresponding to a low load regime. In this latter, the traffic steering at a common RAN NSSMF is not sufficient for the URLLC traffic, and a large control delay worsens the performance. Only a systematic redundant replication is able to achieve the target performance in this low load regime, and this approach can be implemented at the NSMF level, without the need for tight cooperation. Note that an NSMF proportional traffic steering based on the average capacities has the worst performance in the low load regime, but outperforms the redundancy case in high load regime as it avoids overloading.

B. Coexistence of eMBB and URLLC Slices

We now move to a setting where URLLC and eMBB slices share the same resources, i.e. we use the overall bandwidth without reserving a fixed band for URLLC traffic. The overall bandwidth of BS1 and BS2 is 10 MHz and 20 MHz, respectively. Our objective is twofold. First, we aim at studying the impact of eMBB and URLLC slice resource multiplexing on the URLLC performance, and second, we aim at reinspecting the URLLC slice function placement impact. We gradually increase the URLLC traffic while maintaining the eMBB load set to 70% of the cell's capacity.

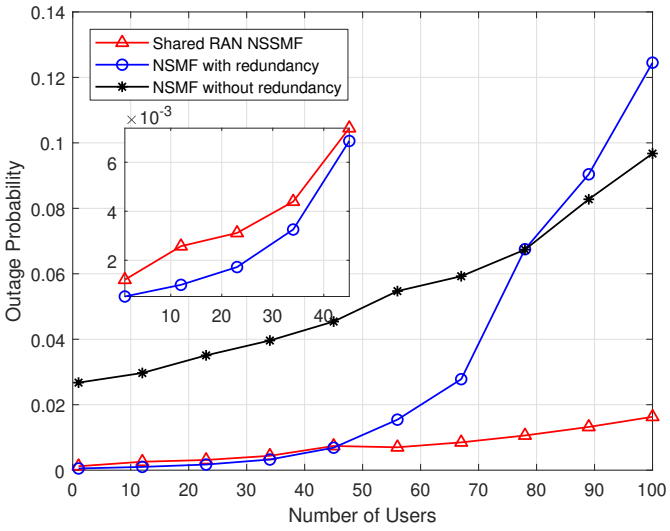

Fig. 5: Outage probability in the case of coexistence of eMBB and URLLC packets without bandwidth reservation.

In Figure 5, we plot the outage probability for URLLC packets. We first remark that the outage probability has higher values compared to the previous case (separated URLLC/eMBB) since the URLLC packets compete with large eMBB packets. Ultra-reliability is thus very difficult to achieve when there is no strict resource reservation for URLLC traffic. We now have a deeper look at the performance of the different URLLC slice management policies. The performance trend is similar to that of the case of resource reservation. In particular, blind redundancy degrades the performance for high loads but is essential for achieving high reliability. Indeed, even if it increases the load, redundancy increases the chance to have a link with no eMBB packet ahead in line. However, tight coordination at the RAN NSSMF level offers altogether good results but still outperformed in low load regimes by packet duplication. In this configuration, the proportional policy offers the worst performance because it is a long term strategy that does not consider the evolution of traffic over time.

V. Concluding Remarks

In this article, we explored the concept of network slicing with the objective of achieving seamless resource allocation at the RAN level while enabling multi-tenancy. We identified the different actors in RAN resource allocation while shedding light on their ownership of the different slice management functions. We studied various options for the placement of intelligent entities involved in resource allocation and traffic steering decisions while focusing on the challenging use case of URLLC traffic in the uplink. Our results show that, regardless of the architectural solutions, redundant scheduling is essential for achieving ultra-reliability.

As of future work, we aim at extending our study to other forms of redundancy, e.g. time and frequency packet replication, in addition to the spatial replication considered in this paper, and explore the feasibility and effectiveness of these schemes for different slicing architectural options.

Acknowledgements

This work is supported by MAESTRO-5G project funded by the French Agence Nationale de la Recherche (ref. ANR-18-CE25-0012).

References

[1] 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, 2017.

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

[3] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

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

[5] 3GPP, "Summary of rel-15 work items," *3GPP TR 21.915 v15.0.0, Tech. Rep.*, September 2019.

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

[7] 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.

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

[9] 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.

[10] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant mobile networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, Apr. 2019. [Online]. Available: https://doi.org/10.1109/TNET.2019.2895378

[11] P. Caballero, A. Banchs, G. De Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6419–6432, 2018.

[12] 3GPP, "Access to the 3GPP 5G core network (5GCN) via non-3GPP access networks (N3AN)," *3GPP TS 24.502 V16.2.0, Tech. Spec.*, December 2019.

[13] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124–130, 2018.

[14] ITU-R, "Minimum requirements related to technical performance for IMT-2020 radio interface (s)," ITU, Tech. Rep., Nov. 2017, Tech. Rep., 2017.