Anticipatory slice resource reservation for 5G vehicular URLLC based on radio statistics

Nathalie Naddeh, Sana Ben Jemaa, Salah Eddine Elayoubi, Tijani Chahed

Anticipatory Slice Resource Reservation for 5G Vehicular URLLC Based on Radio Statistics

Nathalie NADDEH[1,3], Sana BEN JEMAA[1], Salah Eddine ELAYOUBI[2], Tijani CHAHED[3]
[1]Orange Labs; Chatillon, France
[2]Université Paris Saclay; CentraleSupelec; L2S, CNRS, Gif-Sur-Yvette, France
[3]Institut Polytechnique de Paris; Telecom SudParis; Palaiseau, France

***Abstract*—In this paper, we consider resource allocation for vehicular safety traffic. In 5G, this traffic is carried by using the Ultra-Reliable and Low-Latency Communications (URLLC) service as it needs stringent Quality of Service (QoS) requirements in terms of latency and reliability. Since URLLC services may require specific numerology and/or channel access and retransmission strategies, network slicing has been proposed as a solution for QoS requirements and its coexistence with other services such as enhanced Mobile Broad-Band (eMBB). In order to accommodate URLLC traffic, one can opt for static resource reservation, however this is not optimal as it does not follow the real URLLC traffic present in the cell and can impact negatively eMBB traffic. Reactive, on-demand resource reservation is not feasible either as it requires reconfiguration which introduces extra delay that makes it prohibitive to meet URLLC delay requirements. This paper proposes proactive resource reservation schemes that anticipate slice demand. Resource reservation is computed per gNodeB based on the expected traffic and radio conditions. We show how field measurements and trajectory predictions can be used to achieve URLLC objectives with low impact on eMBB performance.**

***Index Terms*—5G, RAN Slicing, Vehicular URLLC, Resource Reservation, Proactive Reservation, MCS Distribution, Dimensioning.**

I. Introduction

A. Context

Vehicle-to-everything (V2X) technologies have been introduced in the Fifth Generation New Radio (5G-NR) under the Ultra-Reliable Low Latency Communication (URLLC) vertical, to cover various applications, such as autonomous driving, vehicle platooning, mission-critical on-board applications like connected ambulances, etc [1]. Their coexistence with other services, especially enhanced Mobile Broad-Band (eMBB) under the same infrastructure is managed thanks to the network slicing paradigm. A slice is a collection of network resources selected to meet Quality of Service (QoS) requirements of a specific service.

Depending on the use case, for instance Remote Driving, the maximum end-to-end latency is set to 5ms and reliability to 99.999% [2] [3]. This requires specific numerology with a small Transmission Time Interval (sTTI) and Doppler-proof design, robust channel coding and matched channel access with replication in contention-based access [4] [5]. This numerology shall be selected based on an algorithm that takes as input the average Signal to Noise Ratio (SNR), the Doppler spread and the delay spread [6]. One or more specific slices must be dedicated to vehicular services, with an appropriately configured numerology in a given bandwidth range.

When provisioning a slice to the URLLC service, 5G-NR defined several enabling technologies such as mini-slot, grant-free and semi-persistent scheduling [7]. In addition to these general URLLC features, when it comes to vehicular services, specific features for coping with high mobility are also provided, with a larger sub-carrier spacing as described in [4]. The resources allocated for the existing slices (V2X and eMBB in this case) must then be configured according to customized numerology. However, as vehicle traffic increases in a Next Generation Node B (gNodeB), more resources need to be added to the corresponding slice and reconfigured before being allocated to this traffic. Re-configuring slices introduces extra delay, on the order of 80 to 100ms [8], making it impossible to meet URLLC stringent QoS requirements. Therefore, proactive reservation schemes that anticipate resource needs and reconfigure the slice before vehicle arrival are required.

B. Related Works

Several works addressed the efficient multiplexing of URLLC/V2X and eMBB slices. The resource allocation algorithm presented in [9] aims to maximize resource utilization and increase eMBB throughput but fails to achieve the required V2X latency and outage probability. The authors in [10] propose a priority-based resource reservation mechanism that aims to reduce URLLC delay and increase its reliability in the presence of eMBB, but their solution does not achieve the 99.999% reliability target. The authors in [11] propose a slice-aware Radio Access Network (RAN) resource allocation for multiplexing eMBB and URLLC users with service isolation. They formulate the problem as an Adaptive Modulation and Coding optimization problem. They aim to maximize the sum-rate of the network but target a Block Error Rate (BLER) of 0.001 which does not meet URLLC reliability constraint.

Other works use scheduling by puncturing, which consists in overlapping URLLC packets on eMBB transmissions. In [12] for instance, the authors present a joint URLLC and eMBB scheduling algorithm where they try to satisfy URLLC demands while maximizing the utility function of the eMBB service, which includes the eMBB throughput degradation. In [13], the puncturing mechanism is used for eMBB and

Low Latency Communication (LLC) traffic aiming to hit low latency target, with the introduction of new recovery mechanisms to minimize eMBB packet loss. The authors in [14] propose an optimization-aided Deep Reinforcement Learning (DRL) approach to improve URLLC and eMBB reliability. Their results satisfy the eMBB requirements but without tackling the URLLC latency constraint and with high URLLC packet outage. These works however do not take into account the slice reconfiguration issue when using different numerology, which we do in the present work.

C. Contributions

To address the reconfiguration delay problem mentioned above and guarantee the URLLC requirements without drastically affecting the eMBB performance, we propose in this paper two proactive resource reservation techniques. The first one is based on locating vehicles at the gNodeB level and proactively reserving resources for them on the target and neighboring cells. The second scheme is based on more refined localization and trajectory anticipation, it pre-allocates resources only to the target and next gNodeB on the trajectory. We compare these schemes to reactive reservation (with reconfiguration delay), and the one based on a maximum static reservation on all cells. In all cases, the eMBB traffic is allocated the left-over capacity.

Preemptive priority is traditionally considered as a solution for ensuring URLLC QoS without resource reservation. This is only possible if the resources for eMBB and URLLC slices are configured with the same numerology (Sub-Carrier Spacing, Cyclic Prefix, channel access) but a different mini-slot size. For services that require specific numerology, preemption is not possible, Radio Resource Control (RRC) reconfiguration is required before eMBB resources can be reused by URLLC.

In our previous work [15], the allocation of Resource Blocks (RBs) for each packet depended on offline simulations. In this work, the number of RBs to be allocated per slice is calculated using a dimensioning method based on the knowledge of the URLLC Modulation and Coding Scheme (MCS) distribution in the different gNodeBs. Based on the Signal to Interference Noise Ration (SINR), the Medium Access Control (MAC) entity returns the MCS that should be adopted. This data is available in the network and is reported to the network management entities.

The rest of this paper is organized as follows: In Section II, we describe the system model and formulate the slice resource allocation problem. Section III details the proposed proactive resource allocation schemes. Section IV develops the performance model to estimate the number of resources to be used and presents the simulation framework. It also compares the performances of the proposed schemes and identifies the best trade-off between the performances of URLLC and eMBB. Section V concludes the paper.

II. System Model

The network consists of a set of K gNodeBs, each one supports two slices for eMBB and V2V/V2I services [16]. Spectral resources should be distributed among these slices to satisfy URLLC Service Level Agreement (SLA) - a packet loss on the order of 10^{-5} and a radio delay[1] less than 3ms- while minimizing the degradation of eMBB throughput.

Network slices are managed at the RAN level by the Network Slice Subnet Management Function (NSSMF), where most of the slice intelligence resides [18]. In particular, the NSSMF should collect and store essential slice quality information, such as observed Channel Quality Indicator (CQI) distribution and QoS, and make intelligent decisions about resource reservation and scheduling schemes.

We develop a dimensioning module that estimates the number of resources that should be reserved for URLLC users for a given traffic demand. We describe how different slices use these resources to serve their traffic and derive corresponding QoS metrics.

A. Radio Model

In 5G NR, the NR numerology can be adjusted in the time and frequency domains [19]. For eMBB, the smallest time allocation is TTI, which is 1ms for 15KHz sub-carrier spacing and includes 14 symbols or Resource Element (RE). For URLLC, the allocation can be performed on a smaller time basis, called sTTI (for short TTI), consisting of 2, 4, or 7 symbols [20]. In the frequency domain, the total bandwidth is divided into sub-carriers. A combination of an sTTI and the 12 sub-carriers is called Resource Block (RB) and corresponds to the smallest radio resource unit assigned to a URLLC user. Note that multiple packets can be scheduled on the same sTTI.

We now determine the number of resource blocks needed for carrying a URLLC packet. We consider a gNodeB where the URLLC slice has to carry packets that belong to different users and thus use different MCS. For an MCS i, let the spectral efficiency be equal to e_i (bit/s/Hz). For an application packet of size a bits, a bandwidth per RB of bHz and a sTTI length of T, the number of physical RBs, R_i, for transmitting an application packet with MCS i is given by:

$$R_i = \lceil \frac{a}{e_i T b} \rceil \tag{1}$$

$\lceil x \rceil$ being the smallest integer larger than or equal to x. While a depends on the application, b and T depend on the radio configuration, and e_i on the chosen MCS.

B. Traffic model

1) URLLC performance model for a fixed MCS: We start with a simple but practical setting where URLLC users always use the same robust MCS that ensures a low BLER. This way, we can avoid additional delays due to channel acquisition, training and MCS adaptation. In this case, with a total available spectrum for URLLC transmission of B_u (in Hz), the total number of RBs is equal to B_u/b (usually

[1]We consider here the radio delay which is a component of the end-to-end delay. For Vehicular URLLC service, the end-to-end maximum delay is typically set between 5 and 20 ms [17]

an integer), and the number of URLLC packets that can be multiplexed per slot is obtained from equation (1) by:

$$K_u(B_u) = \lfloor \frac{B_u/b}{R} \rfloor = \lfloor \frac{B_u/b}{\lceil a/(eTb) \rceil} \rfloor \quad (2)$$

where R is the number of RBs per packet knowing that the considered MCS has a spectral efficiency of e. $\lfloor x \rfloor$ is the largest integer smaller than or equal to x.

URLLC users generate packets sporadically. Let the packet generation process by a URLLC user be Poisson of intensity λ_u packets per second. For a number of active URLLC users equal to N_u, the aggregated packet generation process is Poisson of intensity $N_u\lambda_u$. Packets that are generated during a sTTI, wait until the beginning of the next sTTI to be transmitted. If the accumulated number of packets is less than the maximal URLLC capacity $K_u(B_u)$ in the frequency domain, as determined in equation (2), the remaining packets are stored in a First In First Out (FIFO) queue and served in the next time slots.

Let $M_u(m)$ be the number of packets in the URLLC queue at time slot $m \in [0, \infty]$. This number evolves as follows:

$$M_u(m) = M_u(m-1) - \min\left(K_u(B_u), M_u(m-1)\right) + x_u(m) \quad (3)$$

where $x_u(m)$ is the number of new arrivals during time slot m that is a Poisson random variable of parameter $N_u\lambda_u T$.

For a packet that arrives at time slot m, the worst case radio delay (when it is put at the end of the queue) is computed by:

$$D_u(m, B_u) = T_c + 2kT_{Proc} + (1+2k)T_{tx} + \lfloor \frac{M_u(m)}{K_u(B_u)} \rfloor \quad (4)$$

with

$$T_c = 2 * T_{L_1/L_2} + T_a \quad (5)$$

where T_{L_1/L_2} is the delay of layer 1/layer 2 processing for eNB and UE, T_a is the delay due to alignment, k is the number of retransmissions, T_{Proc} is the delay between scheduling request and uplink grant, and between downlink HARQ and retransmission, and T_{tx} is the transmission time [21].

This delay is averaged over all time slots that have active user arrivals. The average URLLC delay is given by:

$$\bar{D}_u(B_u) = \lim_{m\to\infty} \frac{\sum_{i=1}^{m} D_u(m, B_u)\mathcal{L}_{x_u(m)>0}}{\sum_{i=1}^{m} \mathcal{L}_{x_u(m)>0}} \quad (6)$$

where $\mathcal{L}_c$ is the indicator function that is equal to 1 if condition c is satisfied, and to 0 otherwise.

2) URLLC performance model for heterogeneous MCS: We now consider the case where packets of different users may use different MCS. Let $\mathcal{I}$ be the set of available MCS. When the number of packets waiting to be served is equal to M_u, the system cannot be completely described by M_u, but also by the MCS of each packet. Let $(M_u, \mathbf{I}_u)$ be the system state, with $\mathbf{I} \in \mathcal{I}^{M_u}$ a vector of length M_u, whose k-th element $(I(k))$ is the MCS index for the k-th packet in the queue (the packet at the head of the queue being numbered 1).

Knowing the queue state $(M_u(m), \mathbf{I}_u(m))$ at time slot m, the gNodeB serves in slot m the maximum number of packets $K(m, B_u) \in [1, M_u(m)]$ such that:

$$\sum_{k=1}^{K(m,B_u)} R_{I(k)} \leq \frac{B_u}{b} \quad (7)$$

Constraint (7) ensures that the consumed resources are limited by the amount of reserved URLLC RBs (B_u/b).

At each time slot, the scheduler serves the first $K(m)$ packets and adds the new arriving packets, whose number is $x_u(m)$ as in equation (3), which becomes:

$$M_u(m) = M_u(m-1) - K(m, B_u) + x_u(m) \quad (8)$$

The indices of the new packets are also added to $\mathbf{I}_u$.

3) eMBB QoS model: As of the eMBB performance, let the total available spectrum in the gNodeB be equal to B Hz and the reserved bandwidth for URLLC users at time t be given by $B_u(t)$. The remaining resources $B - B_u(t)$ are shared among the active eMBB users. At time t, let $n(t)$ be the number of active eMBB users, and $e_i(t)$ be the spectral efficiency of the MCS selected by eMBBB user i, the instantaneous throughput for user i is then given by:

$$T_i(t) = \frac{(B - B_u(t))e_i(t)}{n(t)} \quad (9)$$

III. Vehicular slice resource reservation schemes

Based on the observation of degraded vehicular URLLC performance due to the slice reconfiguration delay (shown in Section IV-A, Figure 5), we propose two algorithms which anticipate the resource allocation for the vehicular slice to ensure the required QoS of the vehicular URLLC users.

A. Proactive reservation on neighboring gNodeBs

In this scheme, without prior knowledge of the users' trajectory, we suppose that when a URLLC user arrives at a gNodeB, it can move to any of the neighboring gNodeBs. Thus, appropriate resource reservation is performed on all neighboring gNodeBs so that the QoS of the URLLC vehicle is guaranteed wherever it moves. This is achieved by pre-reserving resources, depending on the expected load at the neighboring gNodeBs. The impact on eMBB throughput can be significant since we reserve for all neighbors that may or may not host URLLC users.

B. Proactive reservation on vehicle trajectories

This second scheme exploits the possibility that the network can anticipate the vehicular URLLC UEs trajectory. When a user arrives at a gNodeB, a refined reservation is performed only at the next gNodeB in the trajectory, without reserving resources on all its neighbors. The impact on the performance of eMBB users is lower than in the previous scheme.

In order to compute the amount of resources that we need to reserve for URLLC users in these two schemes, we develop a dimensioning module that will be described in the following paragraph.

C. Integrating the MCS distribution estimation

We propose to adjust the reservation based on the URLLC MCS distribution for each gNodeB in the network. When we combine this MCS distribution with the predicted number of URLLC users in a gNodeB, we can give an estimation of the required resource reservation for vehicular URLLC users, using the model of section II-B2. This scheme can be integrated into a management module in a real network using two approaches:

- Slow dynamics: a centralized management entity takes as input the MCS distribution of the gNodeBs and the estimated traffic during fixed amount of time. In return, it gives the reservation rate for a specific slice. In this case, the reservation is on the time scale corresponding to the users' mobility dynamics.
- Fast dynamics: a distributed management entity (on each gNodeB) takes as input the MCS distribution of the gNodeBs and the estimated traffic after each scheduling cycle. The time scale corresponding to this operation goes down to the schedulers' assigned TTI.

In our system model, we choose the slow dynamics approach so that the NSSMF is the management entity that gets radio statistics and gives back the resource reservation per slice. Figure 1 illustrates the architecture for implementing the proposed scheme. Within the NSSMF, two modules allow the dynamic management of the slices. First, an MCS distribution module allows building a per-gNodeB MCS distribution. Second, we use this distribution as input for the resource dimensioning module. This latter takes as input the estimated traffic (number of URLLC users per gNodeB) and computes the needed amount of resources to be reserved for the URLLC slice in each of the gNodeBs. The system applies the new configuration, dynamically changing depending on the NSSMF updates. These configurations are eventually used to schedule the users.

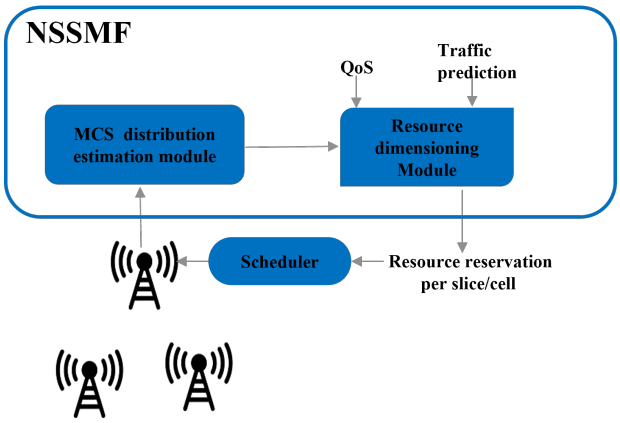

Fig. 1. Integration of the proposed resource dimensioning module.

1) MCS distribution estimation module: The first step in our proposed scheme is to extract the MCS distribution for URLLC slice from field measurements. This is achieved by implementing an MCS distribution estimation module that collects the MCS from user measurement reports on average during the simulation time and associates them to the gNodeB and slice IDs for constructing a per-gNodeB MCS distribution for the URLLC slice.

2) Resource dimensioning module: Let $\mathcal{M}_k = \{p_1^{(k)}, ..., p_{|\mathcal{I}|}^{(k)}\}$ be the MCS distribution extracted from the network as mentioned in the previous section; $p_i^{(k)}$ being the probability of having MCS $i \in \mathcal{I}$ in gNodeB k. The resource dimensioning module associates this distribution with the traffic intensity (in packets/msec) to compute the number of resources to reserve for the URLLC slice in order to achieve the target QoS. We recall that the QoS is expressed as the percentage of correctly received packets within the delay constraint. We implement the following optimization problem:

$$\min B_u \tag{10}$$

subject to the constraint:

$$Pr[D_u(m, B_u) > T_u] \leq \epsilon \tag{11}$$

where B_u is the total available resources for URLLC, $D_u(m, B_u)$ is the per-packet delay of equations (4, 8). T_u is the delay constraint and ϵ is a small positive number.

We solve this stochastic optimization problem using Monte-Carlo simulations. In particular, packets arrive at gNodeB k buffer following a Poisson process and have an MCS that is chosen following distribution $\mathcal{M}_k$. Their number evolves with time following equation (3). The packet's delay is then calculated, leading to the outage probability in (11). Note that the packet arrival rate depends on the predicted number of users in the gNodeB.

Once we obtain the packet loss for known B_u, we search for the lowest resource reservation that achieves the packet loss constraint. We show in Section IV that the MCS distributions can vary from a gNodeB to another depending on the URLLC users' trajectory, and how it affects the required resource reservation (see Figure 3,4).

IV. Performance Evaluation

In our numerical experiments, we implement our NSSMF proposal described in the previous section and illustrated in Figure 1, along with the above-mentioned resource reservation schemes. We develop a 5G network simulator which consists of 13 gNodeBs forming a three-sectored deployment with *500* meters inter-site distance, in compliance with the Third Generation Partnership Project (3GPP) urban macro deployment [22], with 20 MHz bandwidth. We implement network slicing in the NSSMF entity for all gNodeBs. Each gNodeB has two slices: URLLC and eMBB. The slice is created with the following properties: Slice Service Type (SST) [23], label, number of connected users, radio resource percentage, maximum delay, and average throughput. Figure

2 illustrates the network created by the simulator, showing eMBB and URLLC UEs as well as URLLC vehicle trajectory.

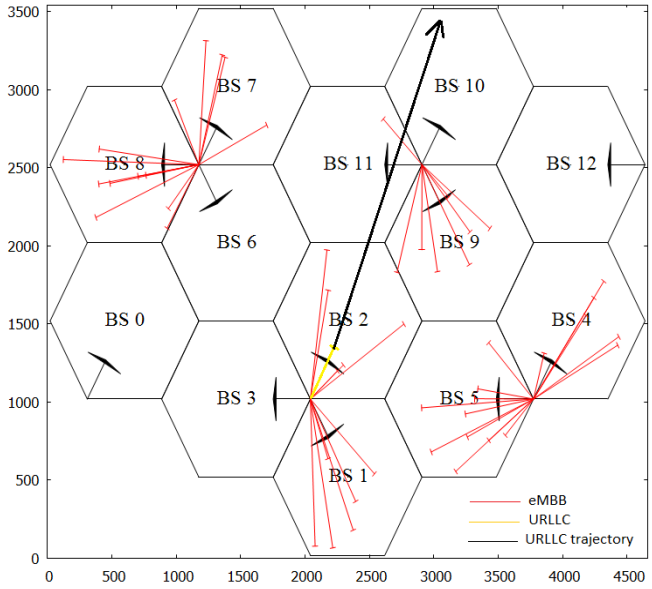

Fig. 2. Urban network with 13 gNodeBs.

The eMBB users arrive in the network following a spatial Poisson process of mean 3.42 [user/sec/gNodeB]. We consider a File Transfer Protocol (FTP) like traffic of fixed file size, 14 Mbits. Once the file is transmitted, the eMBB user leaves the network. Vehicles are created on the roads following a linear Poisson process with different arrival rates depending on the scenario with a mean of 0.395 [Vehicle/sec/km] and move at an average velocity of 50km/h for a total distance of 2.526km. For each vehicle, small URLLC packets of size *96* bits are generated following a Poisson distribution with mean 2 [packets/msec/vehicle]. The vehicles remain active during the simulation time until they leave the network. The simulation and configuration parameters are presented in Table I [5] [24].

TABLE I
SYSTEM PARAMETERS.

Parameters	URLLC	eMBB
Environment	3GPP Urban Macro (UMa)	
Number of gNodeBs	13	
Bandwidth	20 Mhz	
Sub-Carrier-Spacing (SCS)	30 Khz	15 Khz
Number of RBs	51	106
TTI size(ms)	0.143	1
Traffic model	Poisson	
Packet size	96 bits	14Mbits
Speed	50 Km/h	Static
Scheduling granularity	sTTI	TTI

MCS distributions for two different gNodeBs are shown in Figure 3 where we illustrate the probability distribution of the MCS for gNodeB 2 and 10 on one trajectory. We can see that users connected to gNodeB 10 have higher MCS values than those connected to gNodeB 2. This means that users in gNodeB 10 have better radio conditions. The trajectory that is shown in Figure 2 explains this difference, where we can observe that URLLC UEs cross the cell closer to the gNodeB 10 than to gNodeB 2. So we can conclude that gNodeB 2 needs higher RB reservation to compensate degraded radio conditions.

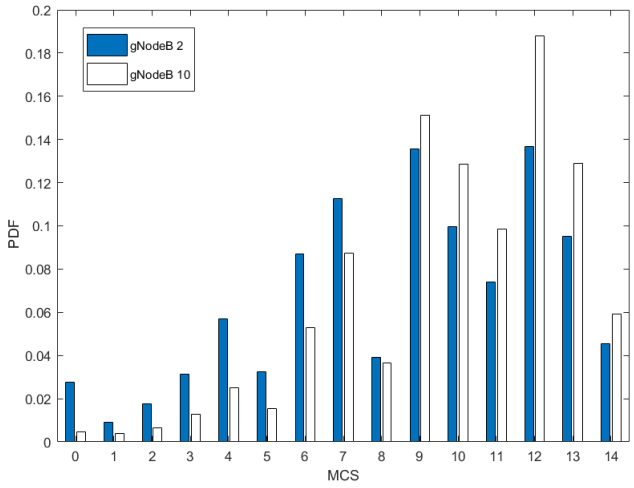

Fig. 3. MCS Distribution.

The resource reservation corresponding to these two gNodeBs is shown in Figure 4. We can see the link between the MCS distribution shown in Figure 3 and the estimated amount of B_u to be reserved for a certain number of URLLC users. We see that gNodeB 2 should reserve a higher quantity of B_u for a specific number of users than gNodeB 10, which matches its radio conditions.

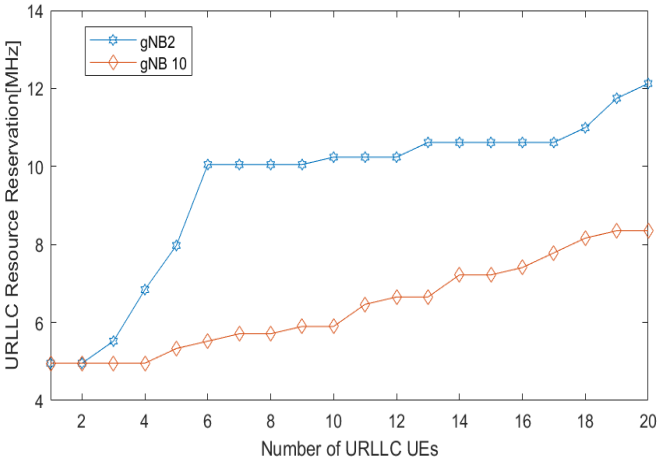

Fig. 4. Resource reservation per number of users in a gNodeB.

The vehicular slice has the following SLA requirements: 10^{-5} of reliability and $5-20$ms of end-to-end latency, which corresponds to radio and back-haul/backbone latency. So depending on the networks' architecture and the services, the

operator can choose the radio latency limit. In our case, we limit the queuing time of the radio latency to $1ms$ after which the packet is considered to be lost. We model the HARQ re-transmission and we take the following assumptions for latency calculation [21]:

- $T_{L1/L2}$ = 1 sTTI,
- T_a = 1 sTTI,
- T_{Proc} = 3 sTTIs
- T_{tx} = 1 sTTI.

We simulate the following resource allocation schemes for URLLC vehicles:

1) *Reactive reservation*, corresponding to the reactive scheme with a reconfiguration delay of 80 ms.
2) *Maximal Static Reservation*, corresponding to a maximal static reservation scheme on all cells, independent of the number of vehicles in the cell. The quantity of reserved resources ensures, on a worst-case basis, that all vehicles would meet their stringent QoS constraints.
3) *Reservation on neighbors*, corresponding to our first proactive reservation scheme on neighboring cells, described in section III-A.
4) *Trajectory Prediction*, corresponding to our second proactive reservation scheme, making use of predicted trajectory, described in section III-B.

A. Reconfiguration impact

As discussed above, RRC procedures, part of the reconfiguration process, take an amount of time that has a negative impact on URLLC delay constraint. We illustrate in Figure 5 the V2X URLLC packet loss during the simulation time due to the handover mechanism. In this simulation, we allocate a minimal amount of resources for URLLC, and we increase the reservation when new vehicles join the gNodeB. With the vehicle's mobility leading to handovers between gNodeBs, we observe peaks of packet losses due to the reconfiguration delay. These peaks vanish for a while until another handover occurs. These peaks can attain a loss of more than 10^{-2} which is unacceptable for V2X URLLC services. One can see that the target of 10^{-5} packet loss is never reached. This motivates the need for our anticipatory reservation scheme proposals.

B. Performance of proactive schemes

In this section, we evaluate the performance of the proactive schemes proposed in this paper and compare them to the reactive and static ones.

We show in Figures 6 and 7 the average URLLC loss probabilities and the average eMBB throughput, respectively for the four reservation schemes, for different URLLC users' arrival rates. For our case, these metrics are the most relevant for the considered slices.

We can see that the reactive scheme has a very high URLLC packet loss, but also the highest eMBB throughput since there is no over-reservation of resources for URLLC users. When an over-reservation of resources is performed in the maximal static scheme, the packet loss of URLLC vehicles reaches very low values at the cost of a very low eMBB throughput. The URLLC packet loss increases when the vehicle arrival rate increases, but it remains below the target of 10^{-5}. However, the eMBB throughput is independent of the URLLC traffic intensity for the static scheme, as the reservation does not depend on the traffic.

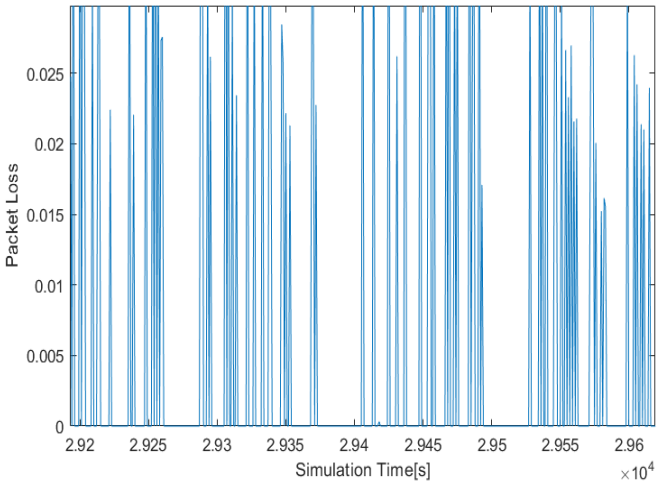

Fig. 5. URLLC packet loss illustration for reactive resource allocation.

When the reservation is performed on the neighbors and anticipated trajectory, we reach acceptable URLLC packet loss values which are below 10^{-5}. However, for the eMBB throughput, we can see the negative impact of reserving extra resources for URLLC on neighbors versus the scheme based on the trajectory. The latter enables a high eMBB throughput, almost equivalent to the reactive scheme, and achieves thus the best balance between URLLC reliability and eMBB throughput.

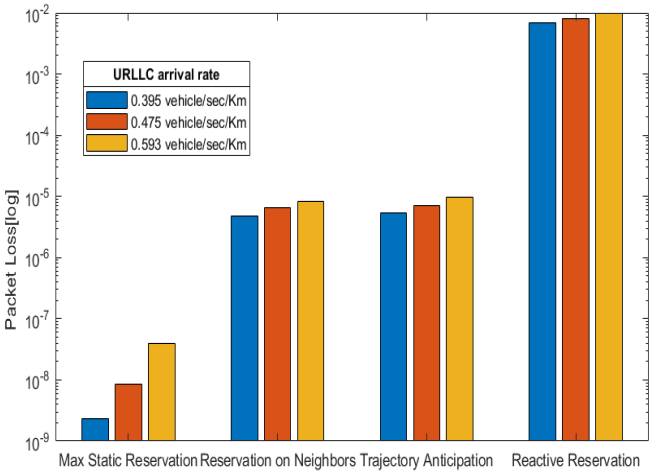

Fig. 6. URLLC arrival rate impact on reliability for the proposed schemes.

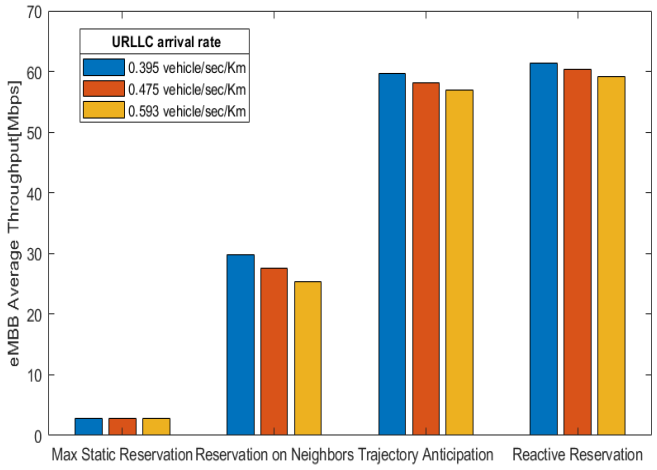

Fig. 7. URLLC arrival rate impact on eMBB throughput.

V. Conclusion

In this paper, we studied 5G network slicing for vehicular URLLC services and focused on the impact of slice reconfiguration delay on its performance. Having observed a URLLC QoS degradation due to this slice reconfiguration delay, we proposed two proactive resource reservation schemes that anticipate slice needs for ensuring URLLC requirements. We developed a per-gNodeB slice dimensioning method to assess the required resources so as to meet vehicle URLLC requirements based on the knowledge of the gNodeB radio condition distribution and traffic intensity.

We studied the performance of the proactive schemes for URLLC and eMBB in a scenario where both slices share the same infrastructure and spectrum, and compared them to the cases with reactive reconfiguration as well as a maximal static resource reservation. We showed that with prior knowledge of the trajectory of the URLLC vehicular UEs, we limited the resource reservation and fulfilled the URLLC requirements while minimizing the impact on eMBB throughput.

In future work, we will extend the joint management of URLLC and eMBB slices to other slice management procedures, including differentiated mobility management and slice-aware load management. We also want to introduce an error to the knowledge of the MCS distribution and analyse the effect on the performance on various RAN metrics.

Acknowledgment

This research work has been funded by the EU Horizon 2020 Program under Grant Agreement No.871249 LOCUS (LOCalization and analytics on-demand embedded in the 5G ecosystem, for Ubiquitous vertical applicationS) project.

References

[1] *Service requirements for V2X services*, 3GPP, TS 22.185, 2020, version 16.0.0 Release 16.

[2] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, 2018.

[3] *5G White Paper*, NGMN, Feb. 2015, version 1.0.

[4] J. Valgas, D. Martín-Sacristán, and J. Monserrat, "5G New Radio Numerologies and their Impact on V2X Communications," *Waves*, 2018.

[5] *Study on NR Vehicle-to-Everything (V2X)*, 3GPP, TS 38.101-1, 2020, version 16.4.0 Release 16.

[6] T. Soni, A. R. Ali, K. Ganesan, and M. Schellmann, "Adaptive numerology—A solution to address the demanding QoS in 5G-V2X," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[7] *Physical layer procedures*, 3GPP, TS 38.213, 2018, version 15.2.0 Release 15.

[8] X. Lin, D. Yu, and H. Wiemann, "A Primer on Bandwidth Parts in 5G New Radio," April 2020.

[9] H. D. R. Albonda and J. Pérez-Romero, "An Efficient RAN Slicing Strategy for a Heterogeneous Network with eMBB and V2X services," *IEEE Access*, vol. 7, pp. 44 771–44 782, 2019.

[10] Y. Chen, L. Cheng, and L. Wang, "Prioritized resource reservation for reducing random access delay in 5G URLLC," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–5.

[11] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN Resource Slicing Mechanism for Multiplexing of eMBB and URLLC Services in OFDMA based 5G Wireless Networks," *IEEE Access*, vol. 8, pp. 45 674–45 688, 2020.

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

[13] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017, pp. 1–6.

[14] M. Alsenwi, N. Tran, M. Bennis, S. Pandey, A. Bairagi, and C. S. Hong, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," *IEEE Transactions on Wireless Communications*, vol. PP, pp. 1–1, 02 2021.

[15] N. Naddeh, S. Ben Jemaa, S. E. Elayoubi, and T. Chahed, "Proactive RAN Resource Reservation for URLLC Vehicular Slice," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021, pp. 1–5.

[16] S. E. Elayoubi, S. Ben 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.

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

[18] A. Chagdali, S. E. Elayoubi, and A. M. Masucci, "Slice Function Placement Impact on the Performance of URLLC with Multi-Connectivity," *Computers*, vol. 10, no. 5, p. 67, 2021.

[19] H. Jang, J. Kim, W. Yoo, and J.-M. Chung, "URLLC Mode Optimal Resource Allocation to Support HARQ in 5G Wireless Networks," *IEEE Access*, vol. 8, pp. 126 797–126 804, 2020.

[20] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[21] *Latency for URLLC*, 3GPP, TDoc R1-1802882, Feb. 2018, version 16.0.0 Release 16.

[22] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects*, 3GPP, TR 36.814, Jan. 2010, version 1.6.0 Release 9.

[23] *System Architecture for the 5G System*, 3GPP, TS 23.501, Dec. 2017, version 15.0.0 Release 15.

[24] "Physical channels and modulation," 3GPP, TS 38.211, 2020, version 16.2.0 Release 16.