On the Scalability of the 5G RAN to Support Advanced V2X Services

M. Carmen Lucas-Estañ[1], Baldomero Coll-Perales[1], Chang-Heng Wang[2], Takayuki Shimizu[2], Sergei Avedisov[2], Takamasa Higuchi[2], Bin Cheng[2], Akihiko Yamamuro[3], Javier Gozalvez[1], Miguel Sepulcre[1], Onur Altintas[2]
[1]Uwicore laboratory, Universidad Miguel Hernandez de Elche, Elche (Alicante), Spain.
[2]InfoTech Labs, Toyota Motor North America R&D, Mountain View, CA, U.S.A.
[3]Toyota Motor Corporation, Japan
[1]{m.lucas, bcoll, j.gozalvez, msepulcre}@umh.es
[2,3]{chang-heng.wang, takayuki.shimizu, sergei.avedisov, takamasa.higuchi, bin.cheng, akihiko.yamamuro, onur.altintas}@toyota.com

***Abstract*— Cellular networks currently support non-safety-critical Vehicle to Everything (V2X) services with relaxed latency and reliability requirements. 5G introduces novel technologies at the radio, transport and core networks that are expected to significantly reduce the latency and increase the flexibility and reliability of cellular networks. This has raised expectations on the possibility for 5G to support advanced V2X applications, including connected and automated applications such as advanced ADAS services, cooperative driving and remote driving. At the radio access network (RAN), 5G introduces the New Radio (NR) interface that incorporates flexible numerologies and new slot formats, channel coding schemes, and radio resource management processes. Previous studies have reported latency values of 5G NR below 2 ms when considering scenarios with limited users in the cell and with unlimited bandwidth. Supporting advanced V2X services using 5G requires a scalable network capable to support a larger number of users without degrading the required service level in scenarios with potentially limited spectrum. This study advances the current state of the art with the evaluation of the scalability of the 5G NR RAN. As a case study, the paper evaluates the capacity of 5G RAN to support the latency and reliability requirements of the cooperative lane change use case as the network load varies. The results show that the capacity of the 5G RAN to support advanced V2X services depends on the system configuration, network load and service requirements. These results call for a careful design, configuration and planning of 5G networks to support V2X services.**

Keywords—5G, NR, New Radio, RAN, Radio Access Network, 5G V2X, V2X, Vehicle to Everything, V2N, V2N2V.

I. INTRODUCTION

Cellular networks can currently support non safety-critical Day-1 vehicle-to-everything (V2X) applications [1] with relaxed latency and reliability requirements. 5G New Radio (NR) improves the latency, reliability and throughput of cellular networks, and offers new opportunities to support advanced V2X applications (also referred to as enhanced V2X or eV2X applications) for connected and automated driving. These opportunities also arise from the introduction of Mobile Edge Computing (MEC) in 5G that increases the computing and storage capabilities at the edge of the network, and therefore facilitates the deployment of V2X services and functionalities closer to the vehicle. This is important because latency is critical for eV2X services.

Previous studies and proof-of-concept trials have shown that 5G NR could support strict latency requirements under certain conditions. For example, the 3GPP's Technical Specification Group Radio Access Network (RAN) evaluated in 3GPP TR 37.910 (v16.1.0) the latency contribution of the radio interface in unloaded conditions and for small IP packets. The results in 3GPP TR 37.910 show that latency values below 2 ms can be achieved in uplink (UL) and downlink (DL) cellular connections in a range of RAN configurations (FDD or TDD frame structure and different numerologies and slot formats). The 3GPP study also shows that high numerologies with shorter symbol time duration can reduce the latency at the RAN to values even lower than 1 ms. Field trials in [2] report sub-2ms over-the-air UL and DL latency values for Vehicle-to-Network-to-Vehicle (V2N2V) connections between trucks operating a platoon. In this case, 5G replaces direct Vehicle-to-Vehicle (V2V) communications between platooning trucks by a connection between trucks through the cellular network. The 5G experimental platform operated at 4.5 GHz, with numerology 2 (60 kHz Subcarrier Spacing –SCS– and 0.25 ms slot duration), self-contained TDD sub-frame structure, Polar coding and 20 MHz bandwidth. These results could open the door for 5G to support safety-critical V2X services that are generally based on direct V2V communications. However, these trials are usually conducted under limited and controlled scenarios, and the question remains on whether 5G networks can scale and support such latency levels to a larger number of users. In this context, this study evaluates the scalability of the 5G NR RAN to support eV2X services. As a case study, this paper evaluates the capacity of 5G RAN to support the latency requirements of cooperative lane change under varying network loads. To this end, we derive a model to estimate the RAN latency under different load conditions and configurations. Our model accounts for the main features of the physical (PHY) and Medium Access Control (MAC) layers that impact the 5G NR RAN latency.

II. 5G RAN LATENCY ESTIMATION

5G RAN has been designed to be flexible and to support a wide range of service requirements. 5G NR defines multiple numerologies and 2 different cyclic prefixes that result in different symbol durations. The numerologies consider different SCS in the frequency domain and slot durations in the time domain as defined in 3GPP TS 38.211 (v16.1.0). The slot duration ranges from 1 ms for numerology 0 with 15 kHz SCS to 0.0625 ms for numerology 4 with 240 kHz SCS. The channel bandwidth is divided into Resource Blocks (RBs) of 12 subcarriers each. 5G NR defines a flexible frame structure and

UMH work was supported in part by the Spanish Ministry of Science and Innovation (MCI), AEI and FEDER funds under Project TEC2017-88612-R, and the Ministry of Universities (IJC2018-036862-I).

the possibility to use slot formats with different number of symbols (full-slots with 14 symbols or mini-slots with 2 to 13 symbols). 5G NR also defines different radio resource management schemes (grant-based and grant-free scheduling) that can be adapted to meet the requirements of different use cases.

To estimate the latency at the 5G RAN, we take into account the configuration of the RAN and the deployment scenario, including the bandwidth, the characteristics of the data traffic, the density of vehicles and the distribution of vehicles. The RAN latency (l_{radio}) can be estimated as:

$$l_{radio}=l_{radio\_UL}+l_{radio\_DL} \quad (1)$$

where $l_{radio\text{-}UL}$ is the UL RAN latency from the User Equipment (UE) to the gNB (or base station) and $l_{radio\text{-}DL}$ is the DL RAN latency from the gNB to the UE.

$l_{radio\text{-}UL}$ and $l_{radio\text{-}DL}$ are computed considering the different factors that contribute to the latency experienced in the transmission of a packet. This includes: 1) the processing delays in the transmitter and receiver ($t_p^{tx\text{-}UE}$ and $t_p^{tx\text{-}gNB}$ when the transmitter is the UE and the gNB respectively, and $t_p^{rx\text{-}UE}$ and $t_p^{rx\text{-}gNB}$ when the receiver is the UE and the gNB respectively); 2) the frame alignment times (t_{fa}); 3) the delay introduced by the scheduling (t_{sch}^{DL} in DL and t_{sch}^{UL} in UL); 4) the waiting time for the allocated resources or RBs (t_w); and 5) the transmission time (t_{tt}). All these factors are illustrated in Fig. 1 for the UL and are explained in more detail afterwards (striped rectangles in Fig. 1 represent the processing of packets in the UE or gNB). $l_{radio\text{-}UL}$ and $l_{radio\text{-}DL}$ are then estimated as expressed in (2) and (3):

$$l_{radio\text{-}UL} = t_p^{tx\text{-}UE} + t_{fa} + t_{sch}^{UL} + t_w + t_{tt} + t_p^{rx\text{-}gNB} \quad (2)$$

$$l_{radio\text{-}DL} = t_p^{tx\text{-}gNB} + t_{fa} + t_{sch}^{DL} + t_w + t_{tt} + t_p^{rx\text{-}UE} \quad (3)$$

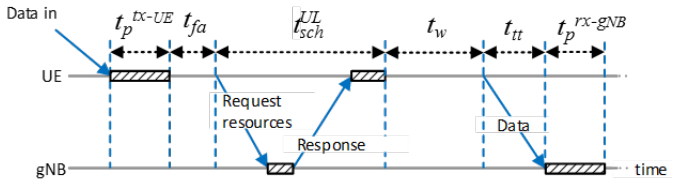


Fig. 1. Latency in the UL transmission of a data packet.

The processing delay in the transmitter ($t_p^{tx\text{-}UE}$ and $t_p^{tx\text{-}gNB}$) represents the time interval between the arrival of data and the generation of a packet. The processing delay in the receiver ($t_p^{rx\text{-}UE}$ and $t_p^{rx\text{-}gNB}$) represents the time interval between the reception of a packet and the decoding of the data. $t_p^{tx\text{-}UE}$, $t_p^{tx\text{-}gNB}$, $t_p^{rx\text{-}UE}$, and $t_p^{rx\text{-}gNB}$ are estimated following 3GPP TR 37.910 as[1]:

$$t_p^{tx\text{-}UE} = t_p^{tx\text{-}gNB} = T_{proc,2}/2 \quad (4)$$

$$t_p^{rx\text{-}UE} = t_p^{rx\text{-}gNB} = T_{proc,1}/2 \quad (5)$$

$T_{proc,1}$ is the UE Physical Downlink Shared Channel (PDSCH) processing procedure time, and $T_{proc,2}$ is the Physical Uplink Shared Channel (PUSCH) preparation procedure time. $T_{proc,1}$ and $T_{proc,2}$ are defined in 3GPP TS 38.214 (v16.1.0) as a function of the UE processing capabilities and the numerology among other factors.

$t_p^{tx\text{-}UE}$ or $t_p^{tx\text{-}gNB}$ might be completed at any time within a slot, and $l_{radio\text{-}UL}$ and $l_{radio\text{-}DL}$ must then also account for the frame alignment time t_{fa} until the start of the next slot (see Fig. 1). t_{fa} is then bounded by the slot duration (1 ms to 0.0625 ms depending on the numerology).

t_{sch}^{UL} and t_{sch}^{DL} depend on the scheduling scheme utilized which should be chosen based on the traffic pattern (e.g. periodic or aperiodic) and service requirements. The dynamic scheduling in 5G NR assigns resources dynamically for each transmission when a packet is generated. In this case, the UE and the gNB must exchange control messages to request/assign the radio resources, and this signaling entails a delay that is detrimental to safety-critical V2X services. We then consider the use of the Semi-Persistent Scheduling (SPS) scheme defined in 5G NR for DL transmissions and a Configured Grant scheduling for UL transmissions. SPS and Configured Grant scheduling pre-assign resources periodically for data transmissions in DL and UL respectively. The scheduling decision is taken (and the resources are pre-assigned) when the UE attaches to the gNB, i.e. before the data packets are generated[2]. Consequently, the UE does not need to request resources to transmit each data packet, and avoids the scheduling signaling delay. In this case, we can assume that both t_{sch}^{UL} and t_{sch}^{DL} are zero which would benefit safety-critical V2X services. Resources are pre-assigned periodically to UEs, and the periodicity between allocated resources is set equal to the transmission rate of data packets.

After t_{sch}^{UL} or t_{sch}^{DL}, the waiting time t_w accounts for the delay until the slot with RBs assigned to the UE is available for data transmission (see Fig. 1). t_w depends on: 1) the size of the packet and the number of RBs necessary to transmit a packet, and 2) the number of free RBs at each slot. The number of RBs necessary to transmit a packet depends on the utilized modulation and coding scheme (MCS) and the number of MIMO transmission layers used. We consider the use of the MCSs defined in 3GPP TS 38.214. Vehicles adapt the MCS based on the Channel Quality Indicator (CQI) table 3 in 3GPP TS 38.214 in order to achieve a Block Error Rate (BLER) target equal to 10^{-5}. The CQI is estimated as a function of the distance between the vehicle and the serving gNB that depends on the distribution of vehicles in the scenario. To compute t_w, we must also estimate the number of free RBs per slot. This number depends on the total number of RBs and the density of vehicles in the scenario. We compute the total number of RBs per slot considering a FDD frame structure. This number is a function of the bandwidth and the numerology as shown in 3GPP TS 38.104 (v16.2.0). From the total available RBs, we identify those that are used by control channels and PHY signals in 5G NR. To this aim, we consider the configuration of the control channels and PHY signals for 5G NR given in Annex A of [3] [3]. This configuration results in that the control channels and PHY

[1] Same processing capabilities are assumed for UEs and the gNB only for evaluation purposes.

[2] There are two types of Configured Grant scheduling in 5G NR. With type 1, the pre-assigned resources are permanently active. With type 2, the pre-assigned resources can be activated/deactivated along the session. We consider type 1 for UL. SPS in DL is similar to Configured Grant type 2.

[3] For the DL, this configuration includes the transmission of the Synchronization Signals and Physical Broadcast Channel (SS/PBCH), the Physical Downlink Control Channel (PDCCH), Demodulation Reference Signals (DMRS), and Channel Status Information (CSI). For the UL, it includes the Physical Random Access Channel (PRACH), the Physical Uplink Control Channel (PUCCH), DMRS, and Sounding Reference Signals (SRS).

signals require approximately 12.5% and 9.3% of the DL and UL available RBs, respectively (for a bandwidth of 40 MHz). The remaining RBs can be utilized to transmit the data packets. Once we know the total number of RBs available and the number of RBs necessary to transmit the packet, we emulate the resource allocation process to identify the free RBs per slot as a function of the density and then estimate t_w. We consider that the scheduler allocates RBs in the first slot where there are enough free RBs to transmit the packet.

To compute $l_{radio\text{-}UL}$ and $l_{radio\text{-}DL}$, we finally need to calculate the latency experienced in the transmission of the data packet. The transmission time or t_{tt} is equal to the length of the slots used to transmit the data packet (see Fig. 1). t_{tt} depends then on the numerology that is used.

III. Cooperative Lane Change Use Case

3GPP identifies in 3GPP TS 22.186 (v16.2.0) the performance requirements for eV2X services related to connected and automated driving. These services are classified in the following five groups: vehicle platooning, advanced driving, extended sensors, remote driving and vehicle quality of service support. For this study, we consider the cooperative lane change use case that is part of the advanced driving group. This safety-critical use case has been traditionally supported using V2V communications (e.g. [4]). It is then interesting to investigate if this use case could be supported through 5G-based V2N2V communications. Please note that cooperative lane change is analyzed as a case study. The authors in no way suggest cooperative lane changes should or should not be executed using V2N2V.

In the cooperative lane change, vehicles exchange driving intentions with vehicles in the proximity in order to coordinate their trajectories and maneuvers. The format, length and periodicity of the messages to be exchanged is currently under study in 3GPP TS 22.186, ETSI TR 103 578 (v0.0.5) and SAE J3186. While no final decision has been made, these sources suggest a message size between 300 and 600 bytes. 3GPP TS 22.186 defines requirements for the cooperative lane change use case. For high level automation, 3GPP establishes that "the 3GPP network shall support message exchange between UEs with less than 10 ms latency, with 99.99% reliability".

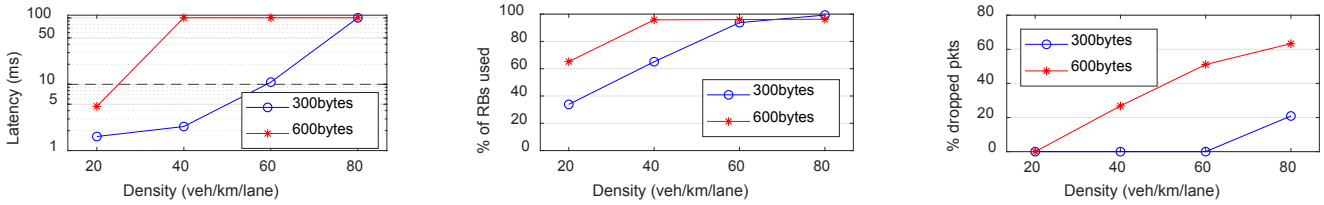
IV. Evaluation

This section evaluates the scalability of the 5G RAN to support the latency requirement of the cooperative lane change use case as the network load increases. To this end, we consider the latency model defined in Section II and numerical evaluations in Matlab. We consider a highway scenario and a single 5G NR cell[4] with 866 m radius as in 3GPP TR 38.913 (v15.0.0). The cell is assigned 40 MHz bandwidth in the Frequency Range 1. The highway consists of 6 lanes per direction and we evaluate vehicle densities equal to 20, 40, 60 and 80 vehicles/km/lane. All vehicles exchange packets of the same size D periodically every 50 ms through the cellular network (D is set to 300 or 600 bytes). It is important that safety-critical V2X services rely on fresh information. As a result, if a new packet is generated and the previous packet has not been transmitted yet, the previous packet is dropped. The transmitted packets include information about planned or desired trajectories. The information transmitted by a vehicle in the UL is transmitted to the neighboring vehicles in the DL. To this aim, we consider the use of the Broadcast/Multicast mode for DL communications even though this mode is not supported in 3GPP Release 15 and Release 16 standards. However, the Broadcast/Multicast transmission mode is highly relevant for efficient support of V2X services, and 3GPP is currently working to include it in Release 17 (3GPP RP-201038). We consider the 5G NR numerology 2 with an SCS equal to 60 kHz with Extended Cyclic Prefix for the transmission of short packets with low latency requirements in UL and DL. The UE processing capability is equal to 2, and $T_{proc,1}$ and $T_{proc,2}$ take values equal to 0.161 and 0.193 ms respectively. We consider full-slot transmissions in UL and DL, the MCSs and CQI tables 3 (i.e. a target BLER equal to10^{-5}) defined in 3GPP TS 38.214, and 2 MIMO transmission layers. SPS and Configured Grant scheduling schemes are used in UL and DL, respectively.

3GPP considers that a data packet is successfully delivered if it is received before the maximum latency requirement defined for the service to be supported. For cooperative lane change with high level of automation, 3GPP establishes a 99.99% reliability requirement with a maximum latency of 10ms. Fig. 2.a plots the 99.99 percentile value of the RAN latency (l_{radio}) for different traffic densities (and hence network loads). The figure includes a dashed horizontal line to identify the 10 ms 3GPP latency requirement. Fig. 2.a shows that the latency increases with the density of vehicles in the cell and the size of packets. The study in 3GPP TR 37.910 showed that a latency value equal to 1.12 ms could be met in unloaded conditions and for small IP packets (only containing the IP header). The latency values experienced as the load increases (Fig. 2.a) significantly overpass 1.12 ms. This clearly highlights that the capacity of 5G networks to support critical V2X services must be evaluated considering the scalability perspective.

Fig. 2.a shows that packets of 300 bytes experience an increase in latency that does not grow linearly with the density and that significantly increases as the density and network load grows. For example, the latency increases by 42% when the vehicle density grows from 20 to 40 veh/km/lane. The increase is equal to 366% and 835% when the density jumps from 40 to 60 and from 60 to 80 veh/km/lane, respectively. This trend is due to the cell congestion. Fig. 2.b depicts the percentage of RBs in the cell as a function of the traffic density. Fig. 2.a and Fig. 2.b show that the latency increases rapidly when the percentage of RBs utilized by UEs in the cell is high and the cell is close to congestion. Fig. 2.b shows that this happens for densities equal to or higher than 60 veh/km/lane when vehicles transmit packets of 300 bytes. The use of RBs increases more rapidly when the packet size is 600 bytes (Fig. 2.b), and cell congestion can occur for densities equal to 40 veh/km/lane. This explains why the 99.99 percentile value of the latency increases rapidly when the packet size is 600 bytes and saturates at 100 ms when the density is equal to or higher than 40 veh/km/lane. We should note that

[4] Analyzing the impact of handovers is not the goal of this paper.



a) 99.99 percentile latency. b) Percentage of utilized RBs. c) Percentage of dropped packets.

Fig. 2. 5G NR RAN performance as a function of the traffic density.

100 ms is the largest possible latency in our study since we drop packets that are not transmitted when a new packet arrives. In this study, packets are periodically generated every 50 ms, so 100 ms is the maximum RAN latency that we can measure at UL+DL. Fig. 2.c plots the percentage of packets dropped as a function of the traffic density. Fig. 2.c shows that 26% of the 600-byte packets are dropped when the density is equal to 40 veh/km/lane. This percentage increases to 51% and 63% when the traffic density increases to 60 and 80 veh/km/lane. On the other hand, 300-byte packets are only dropped in the scenario with the highest density, i.e. 80 veh/km/lane (at this density the percentage of RB utilized increases to 99.3%, see Fig. 2.b). The results in Fig. 2 clearly show that the capacity of the 5G RAN to support the cooperative lane change use case with high levels of automation strongly depends on the network load.

V. Discussion

Previous results highlight the need to consider the scalability of 5G networks when evaluating their capacity to support critical V2X services with low latency requirements. Our preliminary study quantifies the latency at the RAN level and further investigations are necessary to quantify the end-to-end latency of V2N2V communications considering the impact on latency of the 5G transport and core networks. We should note that in this study the RAN does not include the transport network although this is the case in the 3GPP architecture described in 3GPP TS 38.401 (v16.1.0). We can obtain first latency bounds for the transport and core networks using the definitions of the 5G QoS Identifier (5QI) made in 3GPP TS 23.501 (v16.5.1). The 5QI is used to identify/classify the traffic flow of a service and define how it should be handled (scheduling, priority, etc.) to meet its service requirements. For example, the 5QI 86 identifies services that are characterized with a 5 ms maximum one-way (UL or DL) latency (Packet Delay Budget –PDB– in 3GPP TS 23.501) and 99.99% reliability (Packet Error Rate –PER– in 3GPP TS 23.501). The 5QI 86 matches the requirements of the eV2X cooperative lane change use case with high level of automation described in Section III. For this particular 5QI, 3GPP indicates that the one-way delay at the core network should not surpass 2 ms (Core Network Packet Delay Budget –CN PDB–). In this case, the maximum RAN UL+DL latency (radio interface and transport network) should not exceed 6 ms to satisfy the latency requirements of the cooperative lane change with high level of automation.

We should note that the latency experienced in the core network will strongly depend on the specific network deployment. However, 3GPP does not specify what 5G network deployment is appropriate to meet the CN PDB identified per 5QI. The flexibility introduced in 5G offers the possibility for centralized network deployments where the V2X Application Server (AS) and processing are hosted in a central node, or for more distributed ones where the V2X AS and processing are moved to the edge of the network, for example at a MEC node. These distributed deployments can help meet stringent eV2X service requirements at the expense of an increased cost and complexity compared to a centralized network deployment. The processing power at the V2X AS and MEC nodes (when utilized) will also impact the service level provision. The 5G-PPP Architecture Working group has also proposed in [5] the possibility to form "local end-to-end radio paths" when latency is critical, and data is of local nature and does not require network/MEC level processing. In this case, "local end-to-end radio paths" are established among vehicles that communicate via the gNB that purely acts as a forwarder or reflector, and where the transport and core networks do not intervene in forwarding packets. The latency added by the data routing/forwarding functions at the gNB is limited to approximately 200 μs [6].

VI. Conclusions

Previous studies and trials have shown that 5G can support V2X services with low latencies when serving a limited number of users. This paper improves the state-of-the-art by analyzing the scalability of 5G networks and by evaluating their capacity to sustain low latency V2X service level requirements as the network load increases. The study focuses on the latency at the 5G RAN and clearly shows that the capacity of 5G to sustain low latencies at the RAN strongly depends on the network load. Further extensions of this work are planned to refine the modelling at the RAN by including, for example, the effect of retransmissions and the use of mini-slots. Additional steps include the analysis with varying service level requirements and the study of the latency introduced by the core and transport networks.

References

[1] 5GCAR, "Automotive use cases and connectivity challenges, business models and Spectrum related aspects", Deliverable 2.3, July 2019.

[2] K. Serizawa, M. Mikami, K. Moto and H. Yoshino, "Field Trial Activities on 5G NR V2V Direct Communication Towards Application to Truck Platooning", IEEE VTC2019-Fall, Honolulu, HI, USA, 2019, pp. 1-5, doi: 10.1109/VTCFall.2019.8891260.

[3] ITU Radiocommunication Study Groups, "Final Evaluation Report from the 5G Infrastructure Association on IMT-2020 Proposals IMT-2020/14, 15, 16 parts of 17", Document 5D/50-E, Feb. 2020.

[4] B. Lehmann, H. Günther and L. Wolf, "A Generic Approach towards Maneuver Coordination for Automated Vehicles", IEEE ITSC, Maui, HI, 2018, pp. 3333-3339, doi: 10.1109/ITSC.2018.8569442.

[5] 5G PPP Architecture Working Group, "View on 5G Architecture", version 3.0, Feb. 2020.

[6] K. Papagiannaki, et.al., "Measurement and analysis of single-hop delay on an IP backbone network", *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 908-921, Aug. 2003.