Li, Zexian; Shariatmadari, Hamidreza; Singh, Bikramjit; Uusitalo, Mikko

5G URLLC: Design challenges and system concepts

Published in:
International Symposium on Wireless Communication Systems (ISWCS)

DOI:
10.1109/ISWCS.2018.8491078

Published: 12/10/2018

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Li, Z., Shariatmadari, H., Singh, B., & Uusitalo, M. (2018). 5G URLLC: Design challenges and system concepts. In *International Symposium on Wireless Communication Systems (ISWCS)* Article 8491078 (International Symposium on Wireless Communication Systems). IEEE. https://doi.org/10.1109/ISWCS.2018.8491078

5G URLLC: Design Challenges and System Concepts

Zexian Li[1], Hamidreza Shariatmadari[2], Bikramjit Singh[2], Mikko A. Uusitalo[1]
[1]Nokia Bell Labs, Espoo, Finland, E-mail: firstname.lastname@nokia-bell-labs.com
[2]Department of Communications and Networking, Aalto University, Finland, E-mail: firstname.lastname@aalto.fi

Abstract—The upcoming fifth generation (5G) wireless communication system is expected to support a broad range of newly emerging applications on top of the regular cellular mobile broadband services. One of the key usage scenarios in the scope of 5G is ultra-reliable and low-latency communications (URLLC). Among the active researchers from both academy and industry, one common view is that URLLC will play an essential role in providing connectivity for the new services and applications from vertical domains, such as factory automation, autonomous driving and so on. The most important key performance indicators (KPIs) related to URLLC are latency, reliability and availability. In this paper, after brief discussion on the design challenges related to URLLC use cases, we present an overview of the available technology components from 3GPP Rel-15 and potential ones from Rel-16. In addition, coordinated multi-cell resource allocation methods are studied. From the system level simulation results in an urban macro environment, it can be observed that effective multi-cell cooperation, more specifically soft combining, can lead to a significant gain in terms of URLLC capacity.

Keywords—5G, URLLC, latency, reliability, multi-cell/multi-TRP (Tx/Rx Point) coordination;

I. INTRODUCTION

Up to now, the Third Generation Partnership Project (3GPP) has been making good progress in the design of 5G New Radio (NR). And three different service categories have been considered [1]: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC). This is well aligned with the International Telecommunication Union (ITU) requirements for the International Mobile Telecommunications 2020 and beyond (IMT-2020) [2].

Similar to traditional cellular services, eMBB addresses the human-centric use cases for accessing multi-media content, service, and data. mMTC is characterized by a large number of connected devices typically transmitting a relatively low volume of non-delay-sensitive data. URLLC is a communication service for successfully delivering packets with stringent requirements, particularly in terms of availability, latency, and reliability. URLLC will enable supporting the emerging applications and services. Example services include wireless control and automation in industrial factory environments, inter-vehicular communications for improved safety and efficiency, and the tactile internet. It is of importance for 5G especially considering the effective support of verticals which brings new business to the whole telecommunication industry.

With the current Rel-15 work, 3GPP Radio Access Network (RAN) working group aims at providing URLLC service for small data payloads (e.g. 32 bytes) with radio latency of 1 ms (i.e. the latency is measured at the layer 3/2) and with an outage probability of less than 10^{-5} [1]. Regarding to the design target in the upcoming releases for example Rel-16 and beyond, clearly the requirements are even more stringent. The stringent requirements make URLLC a challenging service that entails employing advanced techniques in different parts of the 5G system. To meet such challenging requirements, new technical enablers must be adopted. Such new technical enablers include new numerologies, slot/mini-slot structure, link adaptation, and various diversity techniques for reliability enhancement [3][4][5]. In this paper, after briefly reviewing the challenging key performance indicators (KPIs) from the most demanding use cases, we describe the most critical technology components for supporting URLLC from the aspects of RAN.

The rest of the paper is organized as follows. Section II introduces URLLC requirements. In section III we will discuss the major technical components for latency reduction and reliability enhancement. Section IV introduces one physical layer multi-connectivity and its performance. Finally, conclusions are drawn in Section V.

II. USE CASES AND REQUIREMENTS

In 3GPP, especially within Service and System Aspects (SA) working group, extensive studies have been carried out to understand the most relevant and important use cases and applications from vertical domains [6]. Clearly the most important and well-known use cases requiring ultra-low latency and extremely high reliability are future factory applications, distributed utility grid protection, autonomous driving and so on. The following Table I depicts the relevant KPIs of these use cases. It should be pointed out here only end-to-end latency and reliability are included due to limited space, while there are other key parameters as for example synchronization jitter, network availability, survival time, and user experienced data rate.

Taking industrial applications as one example we can discuss in a bit more detail here in this section. Very likely that industrial applications within a limited area or region, e.g., in a factory, harbor, airport, campus, are the most promising to achieve a positive business case. Considering the most typical use cases of Automation/Motion Control, these are closed-

Table I. Example of low latency and high reliability use cases and their requirements

Scenario	End-to-end latency	Reliability
Discrete automation – motion control	1 ms	99,9999%
Electricity distribution – high voltage)	5 ms	99,9999%
Remote control	5 ms	99,999%
Discrete automation	10 ms	99,99%
Intelligent transport systems – infrastructure backhaul	10 ms	99,9999%
Process automation – remote control	50 ms	99,9999%
Process automation – monitoring	50 ms	99,9%
Electricity distribution – medium voltage	25 ms	99,9%

loop control applications requiring URLLC services, e.g., use of collaborative robots in a factory:

- Latency: from <1 ms to 10 ms.
- Data rate: low as in most cases messages are rather small.
- Reliability: up to BLER $<10^{-9}$.

5G system should be designed not only meeting the end-to-end requirements but also achieving an efficient system deployment. However, network deployment and related optimizations are not in the scope of this paper as our focus is mainly about radio network design to enhance the overall system performance.

III. Technology components To achieve low latency and high reliability

In this section, we discuss various technology components, which are already defined in 3GPP Rel-15 or to be considered in the upcoming releases. These technology components are used to improve system performance especially in terms of latency and reliability, kind of "URLLC tool box". The technology components are classified into two major categories: for latency reduction and for reliability enhancement. However, it should be pointed out that actually these two aspects are not really independent, but rather tightly correlated. For example, the technology components which mainly targeted for low latency could bring more retransmission opportunity which in turn results into more gains in terms of reliability.

A. Low latency

Here we introduce the main technology components from the URLLC tool box for latency reduction.

1) New numerology and transmission time interval (TTI) duration

Table II: OFDM numerologies (normal CP length)

Subcarrier Spacing [kHz]	**15**	**30**	**60**	**120**	**240**
Symbol duration [us]	66.7	33.3	16.7	8.33	4.17
Nominal Normal CP [us]	4.7	2.3	1.2	0.59	0.29
Minimum scheduling interval (symbols)	14	14	14	14	28
Minimum scheduling interval (ms)	1	0.5	0.25	0.125	0.125

In 3GPP, a very flexible frame structure for 5G NR was introduced. The new frame structure can offer different possibilities to shorten the duration of the TTI which is a clear advantage comparing to LTE. For instance, the configurable Subcarrier Spacing (SCS) supports operation in different frequency bands. The 15 kHz SCS corresponds to the baseline configuration, and can be scaled with a factor of 2^N, where N= [0, 1, 2, 3, 4, 5]. The higher SCS, the more symbols can be accommodated in one sub-frame. Table II above illustrates the current agreed Orthogonal Frequency-Division Multiplexing (OFDM) numerologies for 5G NR with normal Cyclic Prefix (CP) length.

On top of this, the number of OFDM symbols per TTI can vary as well. Within NR, the UEs can be scheduled either on slot level of 14 OFDM symbols or on non-slot (a.k.a. mini-slot) level. The length of mini-slot can range from 1 to 13 symbols. Therefore, the reduced TTI length can be achieved by reducing the symbol duration (increasing the SCS) and/or reducing the number of symbols per TTI. For example, a TTI duration of 0.125 ms can be obtained by scheduling users on a slot resolution for the case with 120 kHz subcarrier spacing. Another possibility, more suitable for low frequency bands, is e.g. to use 15 kHz SCS (N=0) and schedule users with a non-slot length of 1-3 symbols (~71-222 μs).

2) Scheduling policy

It is well known that the efficient scheduling algorithms can also reduce the latency. Some of these algorithms are as follows:

- Non-slot based scheduling

Non-slot or mini-slot based scheduling is one key enabler with 5G which is useful in various scenarios especially for low latency transmission. It is envisioned that non-slot based scheduling is essential to fulfill the challenging latency targets especially in lower spectrum.

A mini-slot, the smallest scheduling unit, supports the short transmission duration that is accompanied with the reduced

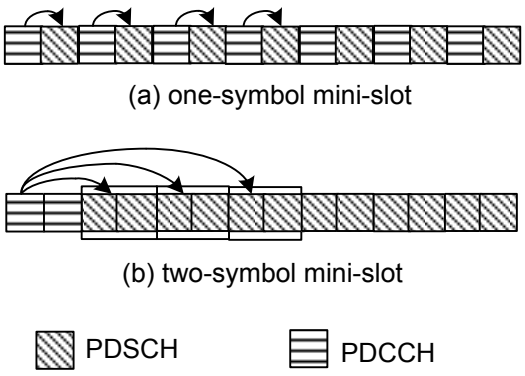

Figure 1: Mini-slot (or non-slot) based DL scheduling

processing time accordingly. The mini-slot length of 2, 4 or 7 symbols is in the recommendation of 3GPP. The supporting of front-load Demodulation Reference Signal (DMRS) in mini-slot enables performing the channel estimation earlier. As shown in Figure 1, at least there are two main approaches for the scheduling of a mini-slot. In addition, across multi-slot/min-slot scheduling is supported as well.

- Efficient URLLC and eMBB multiplexing

How to efficiently multiplexing URLLC and eMBB traffic is one of the key issues to be solved as well. In principle the multiplexing issue covers the scenarios for both inter-UE and intra-UE, also for both DL and UL. For DL, preemptive scheduling [7][8] was specified already. With preemptive scheduling, eMBB traffic is scheduled on all the available radio resources with a long TTI, e.g., 1 ms. When URLLC data arrives at the gNB, it is immediately transmitted to the corresponding UE by overwriting part of an ongoing eMBB transmission. The advantage of this scheme is that the URLLC packets are transmitted without waiting for ongoing scheduled transmissions to be completed. The potential problem with the preemptive scheduling is the degraded decoding performance of eMBB UEs whose transmission is stopped in the middle. To reduce the negative impacts, puncturing indication (PI) was introduced in 3GPP to inform the "victim" UE. The main objective of PI is to tell the UE that part of its transmission has been overwritten. This enables the UE to take this effect into account when decoding the transmission. To be more specific, it knows which part of the transmission is corrupted. Similar concept can be extended to UL as well. For UL multiplexing between eMBB and URLLC services, one scheme which has been proposed and studied in 3GPP RAN1 is the pause-resume scheme (see [9]). With pause-resume scheme, the URLLC UE can take the already allocated resource from eMBB UE.

In addition to inter-UE multiplexing, intra-UE multiplexing between URLLC and eMBB services could be an issue as well. It is possible that the UE could prioritize different logical channels between URLLC and eMBB before data transmissions. However, considering a case that one UE has ongoing UL eMBB transmission while URLLC data arrives, or the UE has an upcoming scheduled UL transmission but does not have sufficient time to prepare URLLC data for this transmission.

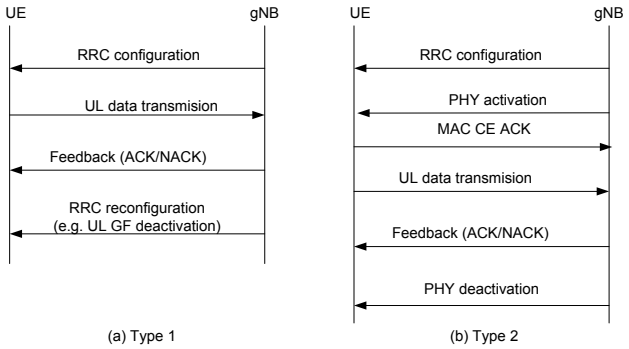

Figure 2 UL grant-free transmission

In such case, one option is that without waiting the URLLC data is sent, using the allocated eMBB resources. This operation is similar to inter-UE puncturing scheduling in DL. Similar scheme as DL PI could help the decoding process at gNB as well.

3) UL grant-free (GF) transmission

For the extremely low latency and reliability requirements, it is desirable to support UL GF transmission scheme (i.e. data transmission without resource request). UL GF transmission can avoid the regular handshake delay: sending the scheduling request and waiting for UL grant allocation. Another advantage is that it can relax the stringent reliability requirements on control channels as well. There are two types of GF configuration schemes supported in 3GPP Rel-15.

For the UL GF type 1, similar as LTE semi-persistent scheduling (SPS), UL data transmission is based on RRC (re-) configuration without any L1 signaling. Potentially SPS scheduling can provide the suitability for deterministic URLLC traffic pattern. This is because the traffic properties can be well matched by appropriate resource configuration. With UL GF type 2 allows additional L1 signaling is introduced. The L1 signaling can be for a fast modification of semi-persistently allocated resources. In such way, it enables the flexibility of UL GF transmission in term of URLLC traffic properties for example packet arrival rate, number of UEs sharing the same resource pool and/or packet size. It is worth to point out that no matter with type 1 or type 2, it is up to gNB configuration to determine whether the resource is exclusive to one UE or not.

B. High reliability

Obviously, the radio link quality affects overall system reliability. Signal to interference plus noise ratio (SINR) is often used to measure the quality of the radio link. The higher the SINR, the lower block error probability, which results in higher reliability and low latency. It is therefore important that a URLLC UE experience SINR above a certain threshold with very high probability. In general, SINR can be increased either by enhancing the signal power, for example with redundancy, diversity and/or to reduce the interference power via interference management. Among the technology components for reliability enhancement, macro diversity will be discussed in Section V.

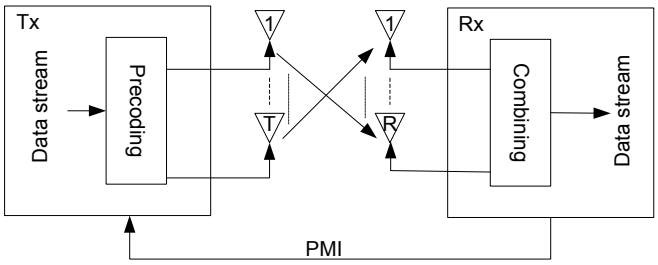

Figure 3: Example of micro-diversity operation with SU single stream transmission

1) Micro-diversity

Micro-diversity refers to the case that having multiple antennas at either the transmitter side or the receiver side or both. One example is shown in Figure 3. For URLLC service, single-user single-stream (i.e. Rank 1) transmission is the most preferred mode due to the design target is to support high reliability.

URLLC link should be operated with at least 2x2 or even more number of antennas. Single-user (SU) single-stream transmission schemes, i.e., maximizing the diversity order of the wireless link should be adopted. As discussed in 3GPP, for the purposes of URLLC performance evaluation, 4x4 was selected to get sufficient diversity orders. With the assumption of independent fading channels among different antenna pairs, high spatial diversity gain can be achieved.

2) High reliable control channels

URLLC service requires the tight latency and high reliability not only for data channel, but also for control channels as discussed in [4], [5]. Below we are discussing various methods for improving the reliability of control channels.

- PDCCH: high aggregation levels

With higher aggregation levels, the control information can be transmitted using excessive resources, by using lower coding rate and/or lower modulation order for reducing the bit/symbol error rates. A high aggregation level of 16 is agreed in 3GPP Rel-15 for URLLC PDCCH transmission.

- Repetition of scheduling information

In our view, increasing the reliability of assignment message can be achieved by including the resource allocation information of current transmission and sub-sequent retransmission. Taking the example shown in the Figure 4, it is assumed that maximal 4 transmissions is allowed and no ACK received during this period. With this assumption, the first assignment message can include the resource allocation information for all the 4 transmissions. In the second slot, the resource assignment information is updated with 3 transmissions only, i.e., from the 2nd to 4th transmissions. Following the same principle, the last scheduling message including the resource information for the last transmission

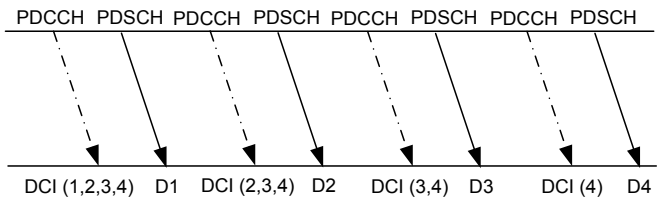

Figure 4: Reliable transmission of DL assignment information

only, i.e., the 4th transmission. With such method, the reliability of the assignment message is increased at the cost of slightly increased signaling overhead. In case UE misses one assignment message, the allocated resource could still be identified with the subsequent assignment message or the previous assignment message. The same operation principle can be of course applied also for scheduling K transmissions for PUSCH as well. The number of repetitions can be flexibly configured by the gNB depending on the reliability target.

- Asymmetric detection of ACK/NACK

As mentioned earlier, protecting the NACK signal is more important than protecting the ACK signal. This is because erroneous NACK detection degrades the communication reliability. On the other hand, wrongly decoding an ACK as a NACK will not result in performance degradation in terms of reliability but on spectral efficiency. This leads to the thought of using enhanced NACK protection by applying the asymmetric signal detection for example [10].

- Adaptive configuring CQI report

The reliability of CQI report itself will bring impacts on the overall reliability as well. When a reported CQI value is decoded wrongly as higher values, it will result in employing a higher MCS and hence reduced the overall communication reliability. One way to enhance the CQI report reliability is to increase the allocated radio resource or decreasing the payload of CQI report. In detail, the potential enhancements can be considered for example:

  - Increased resource for URLLC UE CQI reporting (while keeping the same CQI payload size as eMBB URLLC UEs e.g. 4 bits): with the increased resource, the effective coding rate can be reduced which leads to more reliable CQI decoding at the gNB. This scheme can be supported in Rel-15 already.
  - Another alternative is to define a smaller CQI payload (while keeping the same resources between URLLC UEs and eMBB UEs): in this case the payload of CQI report becomes smaller for URLLC UEs comparing to eMBB UEs. This can lead to a reduced granularity of reporting channel quality. However, with the same amount of resource for CQI reporting, the reliability performance for CQI decoding can be improved.

3) HARQ enhancement

One benefit of the dynamic scheduling scheme is that the network can assign the resources to the UE in a very flexible

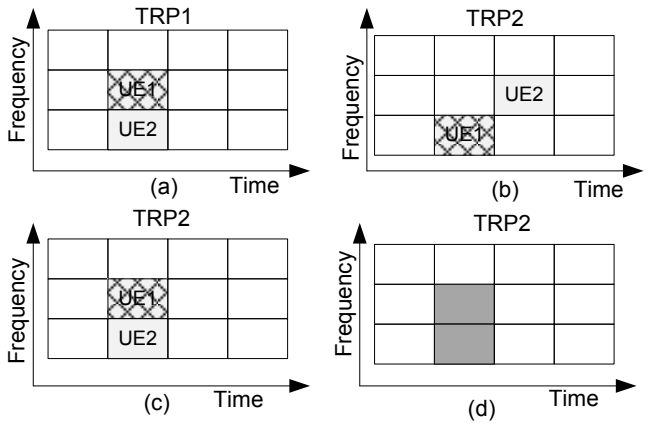

Figure 5: Example of baseline transmission and joint-transmission: (a) baseline; (b) macro-diversity/soft combining; (c) SFN; (d) Narrow-band muting

manner according to the amount of data in the buffer and hence optimize the radio resource utilization. Furthermore, URLLC traffic can be flexibly multiplexed with eMBB. Considering URLLC traffic, one potential concern is the additional latency especially in UL due to the resource request and grant before the UL data transmission. This delay is prolonged by potential HARQ retransmissions also using dynamic scheduling. In order to solve this problem, various enhancements have been discussed to reduce the retransmission latency. One scheme which has been adopted in 3GPP is proactive repetition which can be referred as *K* repetitions as well. With *K* repetition, one UE can get the resource for *K* times transmissions. In case no ACK received for one UL transmission, the UE will automatically transmit the same packet again.

4) Interference mitigation

Mitigating interference by either network-based or UE-based techniques has been identified as a promising complementary solution to improve the SINR. Reducing the received interference from neighboring gNBs or UEs improves SINR. As a rule of thumb, cancelling the strongest or two strongest interferers is usually enough to achieve most of the potential gain. It is expected that interference mitigation will be handled in the coming Rel-16 URLLC work.

IV. Performance Evalution of Macro-diversity

It is well known that various multi-connectivity schemes can provide diversity gain for increasing the reliability. Macroscopic diversity, i.e., data duplication and redundant transmission/reception from multiple cells/TRPs, is also required in order to combat the slow fading effects (or shadowing and/or blocking) and to provide mobility robustness during handovers. In addition, macro-diversity provides benefits in terms of resilience against failures of the cellular infrastructure. In this regard, data duplication at the packet data convergence protocol (PDCP) layer has been agreed for NR in the 3GPP Rel-15. At physical layer, inter-

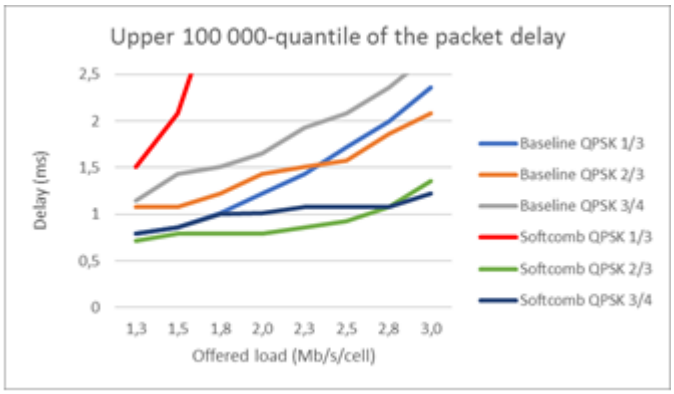

Figure 6: Performance comparison between baseline and the soft combining scheme

cell non-coherent joint transmission is among the promising candidate transmission schemes.

For URLLC, multi-TRP communication can be one of the potential enablers of high reliability. With multiple TRPs, either data packet or control packet or even both can be duplicated among multiple TRPs and sent to the target UEs by multiple TRPs. Different version of the same data packet or the same control information can be received jointly. And UE can potentially combine them in PHY layer. Therefore, the spatial diversity gain can be achieved.

Inter-cell non-coherent joint transmission can be implemented in different ways as discussed in our early paper [11]. Three well-known techniques to increase robustness of a communication link are discussed, namely Single-Frequency Network (SFN) where the same packet is sent with exactly the same resource block from multiple TRPs; narrowband muting with the main target to reduce inter-cell interference and macro-diversity with soft combining. Based on the outcome from [11], it can be observed that in a dense indoor deployment where inter-cell interference is the main reason of degraded performance, inter-cell coordination is a powerful approach to increase the reliability of the transmissions without incurring in longer delays as it is the case of retransmissions.

As extension of our previous work, below we will look at the performance for outdoor scenario defined in 3GPP for URLLC evaluation. Based on our extensive simulations, in case with noise-limited scenario, soft combining is a better candidate. Therefore, below we will focus on the scheme with soft combing. In this scheme, non-coherent transmission of the desired packet is done by the cooperating TRPs, as depicted in Figure 5 (b). The same data packet is sent from multiple TRPs independently. And at UE side, the UE applies soft combining on the received data packets. Such non-coherent transmission can be done independently, such that each TRP performs independent scheduling and with multiple PDCCHs, one per cooperating TRPs.

Figure 6 shows the performance difference between regular single TRP transmission and the studied multiple TRP transmission with soft combing at UE. The 3GPP outdoor simulation environment [8] is adopted and 20% UEs are indoor UEs. From the simulation results, clearly, we can see

that with soft combing, the offer URLLC load is about 2.6 Mbps and in case with regular transmission, the value is about 1.8 Mbps which means more than 40% capacity gain due to the joint transmission from multiple TRPs while keeping the reliability level at 10^{-5} within 1 ms latency.

V. CONCLUSION

In this paper, we have discussed the main challenges related to URLLC services especially from the vertical applications point of view. Clearly, novel technology concepts are necessary in order to fulfill the stringent requirements especially from latency and reliability point of view. Then we introduced the key technology components to reduce latency and increase reliability. Most of the technology components are already specified in Rel-15 including new numerology, DL preemptive scheduling, mini-slot based scheduling, UL grant-free transmission, micro-diversity for reliability, enhanced PDCCH transmission. While there are also features which have high potential in Rel-16 URLLC work as further optimization for example joint multi-TRP transmission/reception, inter-cell interference management etc. Finally, we have presented system-level performance results showing how the macro-diversity can increase the offered URLLC capacity comparing to the regular transmission mode in 3GPP outdoor scenarios without scarifying latency and reliability performance. It can be observed that the targeted URLLC performance can be achieved in the studied scenarios. However, this does not necessarily lead to the same situation in other scenarios, further study/enhancements are needed.

ACKNOWLEDGMENT

The authors would like to thank Kimmo Valkealahti for providing simulation results. Thanks to Harish Viswanathan for valuable comments and thanks to many Nokia colleagues for helpful discussions. This work was supported in part by the Business Finland under the project Wireless for Verticals (WIVE).

REFERENCES

[1] 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies,"
http://www.3gpp.org/DynaReport/38913.htm, Aug. 2017, [Online].

[2] ITU-R M-2083-0, IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, Sept. 2015.

[3] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jantti, "Link Adaptation Design for Ultra-Reliable Communications," in Proc. IEEE International Conference on Communications (ICC), Kuala Lumpur,Malaysia, May 2016.

[4] 3GPP TS 38.300, "NR; Overall description; Stage-2," 2018, [Online].

[5] G. Pocovi et. al., "Achieving Ultra-Reliable Low-Latency Communications: Challenges and Envisioned System Enhancements", IEEE Network, March 2018, Vol. 32, no. 2, Pages: 8 – 15.

[6] TS22.261, Service requirements for the 5G system, v16.3.0, April 2018 [online].

[7] K.I. Pedersen, G. Pocovi, and J. Steiner: "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband", Proc. IEEE Vehicular Technology Conference, September 2017, pp. 1-6.

[8] 3GPP TR 38.802: "Study on New Radio (NR) Access Technology; Physical Layer Aspects".

[9] 3GPP RAN1 document R1-1804618: "On UL multiplexing between eMBB and URLLC".

[10] H. Shariatmadari et al.: "Resource Allocations for Ultra-Reliable Low-Latency Communications", International Journal of Wireless Information Networks, vol. 24, no. 3, September 2017, pp. 317–327.

[11] V. Hytönen, Z. Li, B. Soret, and V. Nurmela: "Coordinated multi-cell resource allocation for 5G ultra-reliable low latency communications", 2017 European Conference on Networks and Communications (EuCNC), June 2017.