5G URLLC Performance Analysis of Dynamic-Point Selection Multi-User Resource Allocation

Ali Karimi[1], Klaus I. Pedersen[1,2], and Preben Mogensen[1,2]
[1]*Wireless Communications Networks (WCN) Section, Department of Electronic Systems, Aalborg University, Denmark.*
[2]*Nokia-Bell Labs, Aalborg, Denmark.*
alk@es.aau.dk

***Abstract*—This paper studies dynamic point selection (DPS) and frequency-selective multi-user scheduling to improve ultra-reliable low-latency communication (URLLC) for the fifth generation New Radio (5G NR) systems. DPS is a special type of multi-channel access scheme enhances the network performance by enabling dynamic transmission point selection on a fast time-scale. The achieved gain from frequency-selective URLLC scheduling is further studied by investigating a low-complexity resource allocation algorithm. Extensive 5G NR system-level simulation results show that DPS achieves 30% improvement of URLLC latency. Our analyses also indicate that for DPS, user-specific clustering with 3-dB power range achieves the major improvement of URLLC latency.**

***Index Terms*—5G new radio, URLLC, Dynamic point selection, Frequency-selective diversity, Packet scheduling.**

I. Introduction

Third generation partnership project (3GPP) has introduced ultra-reliable low-latency communications (URLLC) as a new service class in the fifth generation New Radio (5G NR) [1], [2]. URLLC is envisioned to support a wide range of mission critical applications such as industrial automation, E-health, and vehicular communications, with strict quality of service (QoS) requirements in terms of both reliability (99.999%) and latency (one millisecond) [1], [3]. Lots of studies have addressed challenges that arise from such stringent requirements. As an essential baseline for enabling low-latency communications, the use of short time transmission intervals (TTIs) and flexible frame structure has been investigated in [4]. Dynamic link adaptation and QoS-aware resource allocation of URLLC and enhanced mobile broadband (eMBB) traffic are studied in [5], [6]. The authors of [7], [8] present a survey of the theoretic principles of URLLC and discuss several important enablers for reliable communications. Among the promising solutions enabling URLLC, utilization of massive multiple-input multiple-output (massive MIMO) antennas is investigated in [9]. The use of centralized radio access network (C-RAN) architecture is discussed [10], [11]. The work in [12], [13] present a survey of reliability enhancement of URLLC services through multi-channel access (MCA) solutions.

This paper studies the performance of dynamic point selection (DPS) multi-user resource allocation for URLLC services. DPS is a special case of the MCA family, which provides dynamic transmission point selection on a TTI basis based on channel and load conditions [14]. It is a key feature to mitigate stochastic variations of fading channels for cell-edge users and to enhance the spectral efficiency by enabling fast switching between serving cells.

The concept of DPS has earlier been studied for eMBB traffic in LTE systems to improve the average network capacity [15], [16]. However, given the many differences between LTE and NR, the concept of DPS needs to be revisited to assess its potential performance for URLLC cases. Our starting point is the so-called spectrum efficient DPS, where the users are scheduled by the cell with the highest instantaneous throughput (TP), offering a simple, yet efficient, diversity mechanism. This solution is further extended by pairing it with a latency-aware multi-user diversity resource allocation policy. The proposed solution takes the overhead of control channel transmissions carrying scheduling grants explicitly into account, as well as potential effects of segmentation of the URLLC payloads. The study is conducted for a highly detailed system model in line with 3GPP NR specifications. The model comprises the NR radio access network protocol stack, time-variant URLLC traffic models, a realistic three-dimensional (3D) radio propagation channel, MIMO transmission, dynamic link adaptation, hybrid automatic repeat request (HARQ) retransmission, etc.

Performance results from 3GPP 5G NR compliant system-level simulations are presented to evaluate the performance of the proposed schemes. The results reveal that both DPS and frequency-selective scheduling offer significant latency reduction.

The rest of the paper is organized as follows: Sec-

978-1-7281-2527-5/19/$31.00 ©2019 IEEE

tion II presents an overview of the system model and network deployment. The proposed packet scheduling algorithm is discussed in Section III. Simulation methodology and performance results are presented in Section IV. Finally, Section V concludes the paper.

II. Setting the Scene

A. System Model

We study the downlink (DL) performance for the frequency division duplexing (FDD) mode in line with the 5G NR specifications as outlined in [1], [17]. As in [5], a wide-area urban macro (UMa) scenario of $C = 21$ cells deployed in a three-sectorized manner is assumed. A set of U URLLC user equipments (UEs) are uniformly distributed in the network area. Sporadic traffic is assumed for each UE where bursts of small payloads of 50 bytes arrive at the network following a Poisson point process with an average arrival rate of λ [payload/sec]. The average offered load per cell equals $L = C^{-1} \times U \times \lambda \times 50 \times 8$ [bps/cell].

The UEs are dynamically multiplexed on a shared channel with 20 MHz bandwidth using orthogonal frequency division multiple access (OFDMA) with 30 kHz sub-carrier spacing. A short mini-slot time transmission interval (TTI) of 4 OFDM symbols (≈ 0.143 msec) and physical resource block (PRB) of 12 sub-carriers are assumed.

Both cells and UEs have two transmit/receive antennas. Linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver is assumed at the UEs to suppress the received noise plus interference.

B. Cell connectivity and DPS Procedure

For the baseline (no DPS) scenario, each UE measures cells it can hear and connects to the cell corresponding to the highest received average reference signal received power (RSRP).

Dynamic user-centric clustering is assumed for DPS case. The UE connects to a cluster of maximum Q cells that are within a RSRP power window of W dB as compared to the strongest cell. We denote $\Gamma(u)$ as the set of cells in the cluster for UE u. Channel state information (CSI) measurements are performed periodically for the connected cells and the UE reports channel quality indicator (CQI) of the best cell to the network. Targeting to maximize the instantaneous user TP, UE u reports cell $\hat{c}$ with the highest spectrum efficient metric

$$\hat{c} = \arg \max_{c \in \Gamma(u)} \bar{\mathbf{r}}_c^u, \tag{1}$$

where $\bar{\mathbf{r}}_c^u$ is the estimated full-band TP of the u-th UE served by cell c.

Two types of CQI measurement are performed for the selected serving cell. i) The UE reports one wide-band CQI. ii) One CQI value per a sub-channel of eight PRBs. The CQI is subject to reporting and network processing delay before being applied for the DL transmission decisions.

For each scheduling interval, both the user-specific physical downlink control channel (PDCCH) and the actual data are transmitted on the assigned PRBs. In line with [18], [19], the aggregation level of the PDCCH is dynamically adjusted based on the reported CQI to guarantee low-probability of failure. Dynamic link adaption is adopted for data transmission. The well-known outer-loop link adaptation offset is applied to achieve 1% block error rate of the first data transmission [6], [11]. The UE feeds back a negative acknowledgement (NACK) in case of packet failure and the corresponding HARQ retransmission is scheduled by the network. HARQ Chase-combing is assumed at the UE to increase the quality of received signal by maximum ratio combining (MRC) of the multiple received packets [20].

C. URLLC Latency Components

The DL one-way latency (Υ) of a URLLC payload is defined from the time that the payload arrives at the network until it is successfully received at the UE. If the payload is decoded correctly within the first transmission, the latency equals:

$$\Upsilon = d_{fa,q} + d_{bsp} + d_{tx} + d_{uep}, \tag{2}$$

where $d_{fa,q}$ is the frame alignment and queuing delay. The transmission time is denoted by d_{tx}. The processing time at the base station and user-end are denoted by d_{bsp} and d_{uep}, respectively. The frame alignment delay is a random variable with uniform distribution between zero and one TTI. The queuing delay is the time the packet is buffered before getting scheduled at the physical layer. Depending on the payload size, CQI, and the number of available resources, the transmission time varies between one and multiple TTIs. In line with [21], the processing times are assumed to be constant equal to $d_{bsp} = 2.75$ and $d_{uep} = 4.5$ OFDM symbols, respectively. In case of failure in data transmission, the packet is subject to additional HARQ round-trip-time (HARQ-RTT) retransmission delay(s) (d_{HARQ}). A minimum retransmission delay of $d_{HARQ} = 4$ TTIs is assumed [6].

III. Proposed URLLC Resource Allocation Algorithm

Our target is to maximize the URLLC capacity subject to satisfying both the reliability and latency constraints. The applied radio resource management procedure is as follows. As discussed in Section II-B, each UE dynamically determines the serving cell and periodically reports the corresponding CQI (wide-band/sub-band) to the network. The active UEs with data are allocated resources in each scheduling interval. Building

TABLE I
DEFAULT SIMULATION ASSUMPTIONS.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector BSs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D
Carrier	2 GHz (FDD), 20 MHz carrier bandwidth
PHY numerology	30 kHz sub-carrier spacing configuration. PRB size of 12 sub-carrier (360 kHz).
TTI sizes	0.143 msec (4-symbols mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration. LMMSE-IRC receiver.
CSI	Periodic CSI every 5 msec, with 2 msec latency.
MCS	QPSK to 64 QAM, with same encoding rates as specified for LTE.
Link adaptation	Dynamic MCS with 1% BLER of initial transmission.
HARQ	Asynchronous HARQ with Chase-combining. HARQ-RTT=4 TTIs.
User distribution	2100 URLLC UEs (Average 100 UEs per cell).
Traffic model	FTP3 downlink traffic with payload size of 50 bytes.
Link-to-system (L2S) mapping	Based on MMIB mapping [22].

on [5], a low-complexity resource allocation algorithm is applied to schedule the buffered UEs.

To minimize additional queuing delay, first the HARQ retransmissions are scheduled. For cases with sub-band CQI, the HARQ payloads are scheduled over the set of PRBs with the highest CQI values to enhance the reliability of retransmissions.

Afterwards, pending URLLC payloads are allocated. The time-domain (TD) scheduler selects a subset of UEs closer to the latency deadline which can be fully scheduled on the available resources. The selection metric is expressed as follows

$$\dot{u} = \arg\min_{u \in \Xi(c)} \{\Upsilon^u_{tar} - \Upsilon^u_{cur} \mid R^{uc} \leqslant D^c_{tot}\}, \quad (3)$$

where $\Xi(c)$ is the set of active UEs of c-th cell. The target and current latencies of the u-th UE are denoted by Υ^u_{tar} and Υ^u_{cur}, respectively. The number of available PRBs at cell c and that of required to schedule (both the data and PDCCH) UE u are presented by D^c_{tot} and R^{uc}, respectively. The value of R^{uc} is estimated from the reported wide-band CQI. After selecting UE $\dot{u}$, the scheduler updates the number of available PRBs as $D^c_{tot} = D^c_{tot} - R^{\dot{u}c}$ and search for other schedulable candidate UEs.

For cases with wide-band CQI, the TD selected UEs are randomly allocated over the entire bandwidth. For scenarios with available sub-band CQI, the selected UEs are frequency-domain (FD) multiplexed by allocating resources based on the throughput to average (TTA) metric [5]. PRB p is assigned to UE $\hat{u}$ with the highest TTA metric:

$$\hat{u} = \arg\max_{u \in \mathit{\Pi}} \frac{r^p_u}{\bar{\mathbf{r}}_u}, \quad (4)$$

where $\mathit{\Pi}$ denotes the set of UEs selected by TD scheduler. Variables r^p_u and $\bar{\mathbf{r}}_u$ represent the u-th UE's achievable TP of PRB p and the instantaneous full-bandwidth TP in current TTI.

Finally, the scheduler checks if there are still available resources to schedule additional UE(s). In case of not having enough PRBs to allocate a full payload, only one URLLC payload is segmented and transmitted on the remaining PRBs. To minimize the cost of PDCCH transmission, an UE with the lowest PDCCH overhead (i.e. higher CQI value) is prioritized.

IV. SIMULATION RESULTS

A. Simulation Methodology and Assumptions

The results are generated by running dynamic system-level simulations following 5G NR methodology in 3GPP [1], [17]. Table I summarizes the network settings and simulation parameters. The key performance indicator (KPI) is the one-way latency with 99.999% reliability. The simulation time is set so at least five million packet transmissions are performed, providing reliable results for the 99.999% percentile of the latency [6].

B. Performance Results

Fig. 1 plots the complementary cumulative distribution function (CCDF) of the URLLC latency for different offered loads and scheduling strategies. At eight Mbps offered load, all schemes have similar performance at $1-10^{-5}$ reliability, with latencies between 1.15 to 1.2 msec. At such low offered load, there are only a few active UEs in each scheduling interval. As a consequence, lower levels of inter-cell interference and queuing delay are experienced. Therefore, processing/transmission times, and HARQ-RTT are the dominant factors for the URLLC latency.

Notable latency degradation occurs when increasing the load to 15 Mbps as a consequence of higher queuing delay. Here, the latency performance varies depending on the used scheduling policy. Considering the baseline (no DPS) with wide-band CQI, the outage reliability of 10^{-5} is achieved at 12.2 msec for 15 Mbps offered load. The latency decreases to 2.77 msec by exploiting frequency-selective scheduling. Around 30% improvement is achieved with DPS so the latency is reduced to 1.95 msec. It can be seen that the combination of DPS and frequency selective scheduling results in 85%

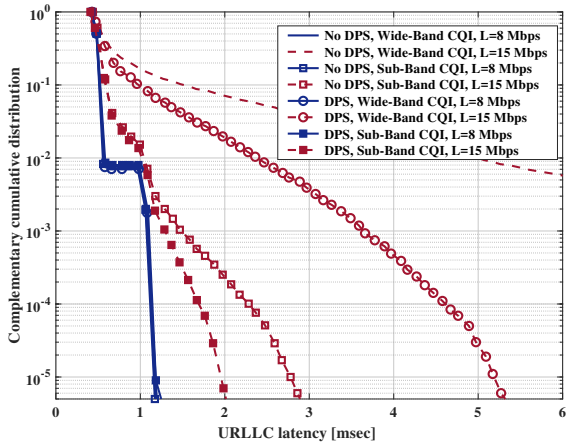

Fig. 1. URLLC latency distribution for different URLLC offered loads and scheduling methods with $Q = 2$ cells $W = 10$ dB.

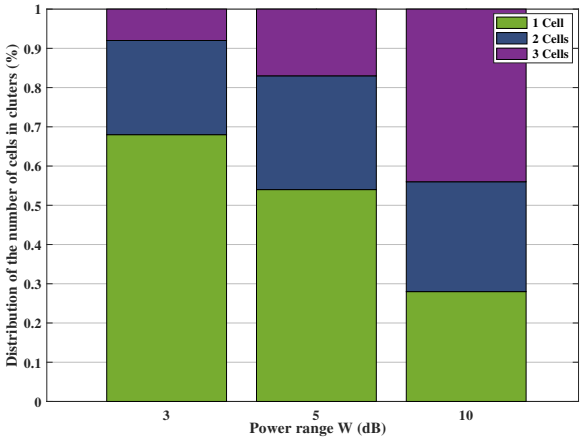

Fig. 3. Distribution of the number of cells in each cluster with respect to different power ranges W.

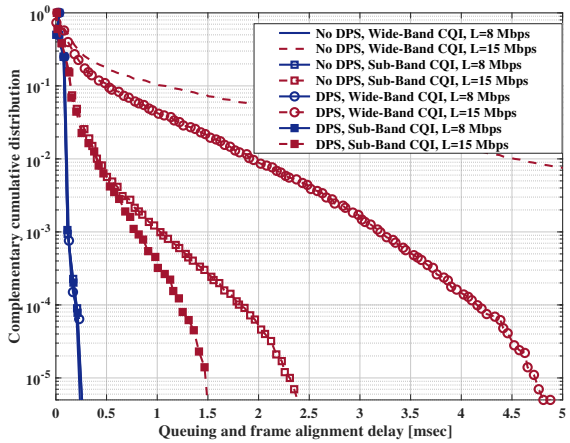

Fig. 2. Queuing and frame alignment delay for different offered loads and scheduling methods with $Q = 2$ cells $W = 10$ dB.

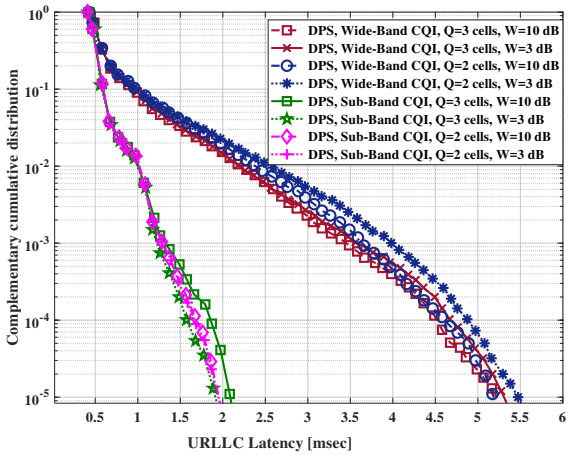

Fig. 4. URLLC latency distribution for different cluster parameters with $L = 15$ Mbps offered load.

latency reduction as compared to the baseline with wide-band CQI. The superior resource allocation by DPS and frequency-selective multiplexing leads to lower number of required PRBs to transmit both data and the PDCCH. As a consequence, the generated inter-cell interference and the queuing delay is decreased.

Fig. 2 depicts the CCDF of the queuing plus frame alignment delay. Although some temporary queuing is observed at low load regimes, the queuing delay has a major impact on the latency degradation when the load increases. Higher packet arrival rate along with the excessive resources required to mitigate inter-cell interference lead to significant negative impact on queuing delay. Fig. 2 shows clear advantages of DPS and frequency-selective scheduling reducing the tail of queuing delay. At 15 Mbps load, 10% of the payloads for the baseline wide-band CQI scenario experience more than one msec queuing delay. With DPS, the number of queued packets decreases by a factor of 2.5. This is further reduced by applying sub-band scheduling to 0.1% and 0.03% for no DPS and DPS cases, respectively.

C. Cluster Variables Analysis

Fig. 3 shows the dynamic cluster size distribution for different values of W, for $Q = 3$. We observe that with $W = 3$ dB, only 32% of the UEs have more than one cell in their cluster, while only 8% of UEs have three cells. As expected, the number of cells in the cluster increases with W. Assuming $W = 10$ dB, in 72% of cases there are at least two cells in clusters while, 44% of UEs have three cells in their cluster sets.

The impact of different cluster parameters is pictured in Fig. 4. As can be seen, the major improvement of DPS is achieved for cases with the power range of $W = 3$ dB, where only 32% of clusters have more than one cells. The results indicate that higher value of the power range does not provide additional latency gain. It

is less probable for the cells with relatively lower signal strength to provide sufficient spectral efficiency. Our observations confirm that DPS affects mainly cell-edge UEs that receive similar signal power from neighbouring cells.

V. Conclusion

We have studied frequency-selective multi-user scheduling and DPS performance of URLLC in 5G NR. Extensive system-level simulations show significant reduced latency of URLLC services at high load scenarios. As an example at 15 Mbps offered load, DPS achieves 30% latency improvement at $1-10^{-5}$ reliability. Exploiting the benefits of both DPS and frequency-selective scheduling offers 85% latency improvement as compared to wide-band CQI scheduling. The results show that DPS is mainly beneficial for cell-edge UEs where major improvement is achieved for dynamic user-specific clustering with the power range of $W = 3$ dB. Future studies should examine load-aware DPS algorithms, impact of non-ideal backhaul, and channel quantization error on the URLLC performance.

Acknowledgement

Part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

References

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 2.0.0, December 2017.

[2] IMT Vision, "Framework and overall objectives of the future development of IMT for 2020 and beyond," International Telecommunication Union (ITU), Document, Radiocommunication Study Groups, February 2015.

[3] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[4] K. I. Pedersen *et al.*, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, March 2016.

[5] A. Karimi *et al.*, "Efficient low-complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," *In Proc. 2019 IEEE 89th Vehicular Technology Conference - VTC2019-Spring*, May, 2019.

[6] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[7] P. Popovski *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[8] 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, October 2018.

[9] T. K. Vu *et al.*, "Ultra-reliable and low-latency communication in mmWave-enabled massive MIMO networks," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2041–2044, Sep. 2017.

[10] A. Karimi *et al.*, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72 253–72 262, 2018.

[11] A. Karimi *et al.*, "Centralized joint cell selection and scheduling for improved URLLC performance," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2018, pp. 1–6.

[12] N. H. Mahmood *et al.*, "Multi-channel access solutions for 5G new radio," *IEEE Wireless Comm. Magazine., 2018, submitted.*

[13] ——, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G New Radio," *CoRR*, vol. abs/1904.07963, 2019. [Online]. Available: http://arxiv.org/abs/1904.07963

[14] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 743–764, Secondquarter 2017.

[15] R. Gupta, S. Kalyanasundaram, and B. Natarajan, "Dynamic point selection schemes for LTE-A networks with load imbalance," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, September 2015, pp. 1–5.

[16] R. Gupta, S. Kalyanasundaram, B. Natarajan, and M. Sen, "Performance analysis of enhanced dynamic point selection CoMP scheme for heterogeneous networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[17] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2016.

[18] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, March 2018.

[19] K. I. Pedersen *et al.*, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *2016 IEEE Global Communications Conference (GLOBECOM)*, December 2016, pp. 1–6.

[20] D. Chase, "Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Transactions on Communications*, vol. 33, no. 5, pp. 385–393, May 1985.

[21] 3GPP Technical Documents R1-1808449, "IMT-2020 self-evaluation: UP latency analysis for FDD and dynamic TDD with UE processing capability 2 (URLLC)," August 2018.

[22] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3766–3778, October 2010.