MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks

Gerardino, Guillermo Andrés Pocovi; Alvarez, Beatriz Soret; Pedersen, Klaus I.; Mogensen, Preben Elgaard

Published in:
2017 IEEE International Conference on Communications Workshops (ICC Workshops)

DOI (link to publication from Publisher):
10.1109/ICCW.2017.7962790

Publication date:
2017

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Gerardino, G. A. P., Alvarez, B. S., Pedersen, K. I., & Mogensen, P. E. (2017). MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks. In *2017 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1005-1010). IEEE (Institute of Electrical and Electronics Engineers). https://doi.org/10.1109/ICCW.2017.7962790

MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks

Guillermo Pocovi[1], Beatriz Soret[2], Klaus I. Pedersen[1,2], Preben Mogensen[1,2]

[1]Department of Electronic Systems, Aalborg University, Denmark,
[2]Nokia - Bell Labs, Aalborg, Denmark
E-mail: gapge@es.aau.dk

***Abstract*—Ultra-reliable low-latency communications (URLLC) entail the transmission of sporadic and small packets, with low latency and very high reliability. Among many potential areas of optimization for URLLC, the problems of large delays during HARQ retransmissions, and inaccurate link adaptation as a consequence of the rapidly-varying interference conditions are studied. The former is addressed by reducing the TTI length and HARQ round-trip time, as compared to what is used in LTE; whereas including low-pass filtered interference information in the CQI report is also proved to have great potential. Extensive system-level simulations of the downlink performance show that the URLLC requirements, i.e. latencies below 1 ms and 99.999% reliability, are achievable at low load scenarios, whereas some performance degradation (1 - 3 ms latency) is experienced at higher loads due to the increased queuing delay and inter-cell interference.**

I. Introduction

Standardization activities towards a fifth generation (5G) New Radio (NR) are gaining big momentum in the 3rd Generation Partnership Project (3GPP) [1]. Ultra-reliable low-latency communication (URLLC) has been agreed as one of the three main use cases, targeting transmission of relatively small payloads with very low latency (<1 ms) and high reliability (99.999%) [2].

The very challenging requirements of URLLC have raised significant attention in academia and industry. For example, the studies in [3], [4] analyse the required improvements on the wireless link. It is shown how a combination of micro- and macroscopic diversity can achieve the signal quality outage probability required for URLLC. In [5], different deployment strategies (e.g. number of cells, frequency reuse pattern) are studied to meet the coverage requirements in a factory automation scenario, while the complementary system-level simulation results are presented in [6]. To achieve low over-the-air transmission delay, reducing the transmission time interval (TTI) is of significant importance [7]. Related to the former, the study in [8] evaluates the downlink latency performance under different TTI durations and load conditions, assuming a mixture of high-priority bursty traffic and best-effort mobile broadband (MBB) traffic. In [9], a link adaptation strategy is presented where the modulation and coding scheme (MCS) is selected according to the target reliability and feedback channel imperfections, assuming that one hybrid automatic repeat request (HARQ) retransmission is allowed.

The applicability of these medium access control (MAC) layer schemes on URLLC is mainly limited by two factors: (i) the relatively large delay that characterizes the HARQ operation, and (ii) inaccurate link adaptation due to the very sporadic URLLC traffic. Motivated by this, we present different enhancements for supporting URLLC in cellular networks. Among others, decreasing the TTI length and HARQ round-trip time (RTT), as compared to LTE, is suggested as a resource-efficient way to improve the latency performance; whereas including time-filtered interference information in the channel quality indicator (CQI) report is proposed to improve the link adaptation accuracy. The benefit of the presented enhancements is evaluated by analysing the downlink latency and reliability performance in a multi-user multi-cell scenario. As it will be shown, the gain provided by each solution depends on the offered traffic load, the inter-cell interference, etc. Given the complexity of the considered problem, the chosen evaluation methodology is dynamic system-level simulations, following the latest URLLC modelling assumptions agreed in 3GPP [10]. Good practice is applied in order to generate trustworthy results.

The rest of the paper is organized as follows: Section II outlines the considered network and traffic model, and the performance metrics. Section III presents the proposed URLLC enhancements. The simulation assumptions are presented in Section IV. Performance results are shown in Section V, followed by concluding remarks in Section VI.

II. Network Model and Performance Metrics

System-level evaluations are a powerful tool to analyse the overall behaviour and performance of cellular systems. Particularly, the 3GPP has highlighted the importance of carrying out highly detailed system-level evaluations in order to analyse the impact of time-varying inter-cell interference, and queuing and scheduling delays, on the URLLC performance [10].

A. Network Layout & Traffic Model

We follow the modelling and assumptions recently discussed in [10], [11]. A macro-cellular network composed of 7 three-sector sites with 500 meter inter-site distance is assumed.

A fixed number of URLLC user equipments (UEs) are uniformly distributed across the network. Unidirectional downlink traffic following the so-called FTP Model 3 (FTP3) is applied. This consists of relatively small packets (typically between 32 and 200 Bytes) that are generated for each UE in the downlink direction following a Poisson arrival process [10].

B. Performance Metrics

In line with [1], the key performance indicator (KPI) is the one-way downlink latency that can be achieved with a $1-10^{-5}$ probability. The latency is measured from the moment a FTP3 packet arrives at the base station until it is successfully received at the UE. This accounts for the queuing delay in the cell, defined as the time elapsed between the arrival of the packet at the base station buffers and the execution of the scheduling decision; frame alignment, i.e. time remaining to the beginning of the next TTI; and transmission delay, including the potential HARQ retransmissions that could occur.

III. URLLC Enablers

A. Low Latency Frame Structure

We adopt the candidate frequency-division duplex frame structure presented in [12]. It consists of a grid of time-frequency resources where users are dynamically multiplexed via orthogonal frequency division multiple access (OFDMA). The time domain is organized into subframes, each containing a set of physical resource blocks (PRB) in the frequency domain. The physical layer numerology follows the recent agreements in 3GPP: 15 kHz sub-carrier spacing (SCS), 14 OFDM symbols (1 ms) subframe, and a PRB size of 12 sub-carriers (180 kHz) as the baseline configuration; although options with 2^N scaling of the SCS, e.g. 30 kHz or 60 kHz, are also allowed [11]. The 3GPP has also agreed on using different TTI durations depending on the user-specific requirements. Apart from scheduling with a 1 ms subframe resolution, smaller scheduling units composed of e.g. 7 OFDM symbols (0.5 ms 'slot') or 2 OFDM symbols (0.143 ms 'mini-slot'), are also considered for 5G [13].

In line with [12], the control channel (CCH) is accommodated within the resources assigned to each user (i.e. in-resource CCH). The CCH contains the scheduling grant indicating the specific time-frequency resource allocation for each user, among other relevant link adaptation parameters required to decode the data. The coding rate of the user-specific CCH is dynamically adapted to the user's channel conditions, following the link-level performance specified in [14]. Note that although scheduling with a mini-slot provides the lowest latency, it has a cost in terms of higher signalling overhead due to the need of sending more frequently CCH information [12].

B. Short HARQ RTT

The HARQ RTT is assumed to scale linearly with the TTI length. Assuming a LTE-alike asynchronous HARQ operation

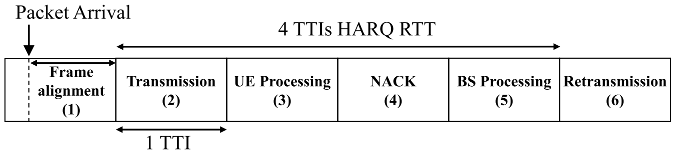

Fig. 1: Diagram of HARQ operation with 4 TTIs RTT.

with a minimum RTT of 8 TTIs, even with a TTI duration of 0.143 ms, the HARQ RTT would not satisfy the 1 ms URLLC latency target (i.e. $8 \cdot 0.143$ ms > 1 ms). Since relying on a single (very conservative) transmission would have a significant cost on the spectral efficiency [9], we study the case where the HARQ RTT is reduced to 4 TTIs in order to allow room for one HARQ retransmission. A diagram of the HARQ procedure with reduced RTT is presented in Fig. 1. We consider a maximum of 1 TTI for frame alignment and for performing the scheduling decision (1). The UE processing time (3) required to decode the initial transmission, and the base station processing (5) of the negative acknowledgement (NACK) are also reduced to 1 TTI[1]. Under such conditions, the maximum latency assuming one HARQ retransmission is reduced to $6 \cdot 0.143$ ms $= 0.86$ ms (excluding the queuing delay). Note that the proposed reduction of the HARQ RTT mainly requires an improvement of the processing capabilities at the UE and base station. A complementary way to relax the processing requirements is by applying the so-called early-feedback techniques [16], which try to predict the outcome of the decoder before the entire decoding finishes.

C. Accurate Link Adaptation

The traffic characteristics of URLLC represent a challenge when attempting to perform accurate link adaptation. Due to the relatively small payloads, a URLLC transmission generally occupies a subset of the available radio resources (i.e. PRBs) within a TTI. This fact, together with the sporadic arrival of packets (as specified in Section II-A), result in a rapidly changing interference pattern. As a result, it becomes difficult to accurately select an appropriate MCS that fulfils a certain block error rate (BLER) constraint. This problem is also well-known from LTE system-level performance analyses in non-fully loaded networks. In such cases, the MCS selection is typically improved by use of outer loop link adaptation (OLLA) mechanisms [17], which "fine tune" the MCS selection according to the received HARQ ACK/NACK feedback messages. These mechanisms are, however, characterized by slow convergence which limits their applicability to URLLC use cases [17].

Our proposal is to modify the UE measurement procedure of the CQI report, by including historical information of the experienced interference. On each TTI n, the UE measures the interference with a certain PRB resolution (a.k.a. sub-band). The interference measurement on the i-th sub-band, $x_i[n]$, is

[1] The processing time would require further reduction if considering uplink-downlink frame misalignment due to the timing advance [15].

filtered with a low-pass first-order Infinite Impulse Response (IIR) filter, resulting in the following smoothed value:

$$s_i[n] = \alpha \cdot x_i[n] + (1-\alpha) \cdot s_i[n-1] \ , \qquad (1)$$

where α is the forgetting factor (FF) of the filter ($0 < \alpha < 1$). The CQI, which is periodically reported to the base station, contains the low-pass filtered interference information together with the latest desired-signal fading information. Note that the latter varies in a much lower time scale and, except for very high UE speeds, it is possible to track the channel variations with relatively high accuracy [15]. The FF α determines how much weight is given to the latest measurement as compared to the previous ones. Based on a heuristic analysis using simulations, it has been found that a FF $\alpha = 0.01$ is beneficial for the latency performance.

D. BLER Optimization

As presented in [8], there is a tradeoff between spectral efficiency and queuing delay. That is, as the system load increases, it is beneficial from a latency point of view to configure the system for high spectral efficiency (rather than for low latency) in order to cope with the non-negligible queuing delay. This can be achieved by adjusting the BLER target for the link adaptation depending on the system load. At low load, conservative transmissions (e.g. 0.1% BLER target) provide the best latency performance, whereas more aggressive transmissions can be allowed at high load, since the reduction of the queuing and scheduling delay compensates for the larger delay in the air interface (due to the occurrence of a larger amount of retransmissions).

IV. Simulation Assumptions

The performance evaluation is based on system-level simulations of a multi-user cellular system. The default simulation assumptions are summarized in Table I. The simulator time-resolution is one OFDM symbol, and it includes explicit modelling of the majority of radio resource management functionalities such as packet scheduling and HARQ. Dynamic link adaptation is applied for both data and the in-resource control channel, which results in varying control overhead depending on the user signal quality and TTI duration (see [8], [12]). Additional overhead from reference signals (RS) is also included. The link adaptation for the data transmissions is based on the periodical frequency-selective CQI report from the URLLC users, using standard OLLA to reach a certain BLER target (0.1% as default). Closed-loop 2x2 single-user single-stream MIMO is assumed for each link and the UE receiver type is minimum mean square error with interference rejection combining (MMSE-IRC).

The network layout, UE distribution and traffic follow the description presented in Section II-A. A set of $N = 210$ URLLC UEs are uniformly distributed in the network, which corresponds to an average of 10 UEs per cell. For each UE, a payload of size $B = 200$ Bytes is generated in the downlink direction following a Poisson distribution with arrival rate λ. The network offered load corresponds therefore to $N \cdot B \cdot \lambda$.

TABLE I: Simulation assumptions

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance [10]
Carrier configuration	10 MHz carrier bandwidth at 2 GHz
PHY numerology	TTI sizes of 0.143 ms, 0.5 ms, and 1 ms; Other numerology settings in line with LTE
Control channel	In-resource control channel scheduling grants with dynamic link adaptation [12]
Data channel MCS	QPSK to 64QAM, with same coding rates as in LTE; 0.1-10% BLER target for first transmissions
RS overhead	4 resource elements per PRB
Antenna configuration	2 x 2 single-user single-stream MIMO with MMSE-IRC receiver
Packet scheduler	Proportional Fair
CSI	LTE-alike CQI and PMI, reported every 5 ms; Interference filtering is applied with FF α
HARQ	Async. HARQ with Chase combining; Max. 6 HARQ retransmissions with 4-8 TTI RTT
RLC	RLC Unacknowledged mode
UE distribution	210 UEs uniformly distributed in the network; 20% indoor, 80% outdoors; 3 km/h UE speed
Traffic model	FTP Model 3 downlink traffic with 200 Bytes payload size
Offered load	1 - 8 Mbps average load per cell

The latency (defined in Section II-B) of each downloaded FTP3 packet is collected and used to form empirical complementary cumulative distribution functions (CCDF). In line with [1], the main KPI is the achievable latency at the 10^{-5} percentile. The latency is analysed both globally and on a per-user basis. The simulation time corresponds to at least 5.000.000 successfully received packets to ensure a reasonable confidence level for the considered performance metrics.

V. Performance Analysis

Next, we present performance results in order to highlight the benefit obtained from the proposed enhancements.

1) Impact of the TTI size: Fig. 2 shows the CCDF of the URLLC latency with TTI sizes corresponding to 2, 7, and 14 OFDM symbols, and an average offered load of 1 Mbps per cell. The system utilization for each configuration, i.e. percentage of PRBs transmitted on average, is shown in the legend. At the 10^{-5} percentile, there is a significant benefit of using short TTI size as it reduces the over-the-air transmission delay, frame alignment, and HARQ retransmission time. The latter particularly impacts the tail of the distribution, since one retransmission typically occurs at the 10^{-5} outage level. This results in a latency of 10 ms for the 1 ms TTI, whereas it is only 3.3 ms for the 0.14 ms TTI. The performance gain of the 0.14 ms TTI as compared to the 1 ms TTI is much smaller (3x shorter latency) than the expected (7x). This is due to

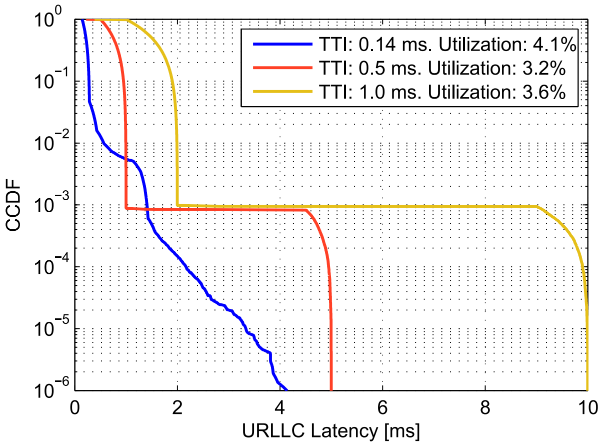

Fig. 2: Latency distribution with different TTI lengths and 1 Mbps offered load. HARQ RTT: 8 TTIs. BLER: 0.1%. $\alpha = 1$.

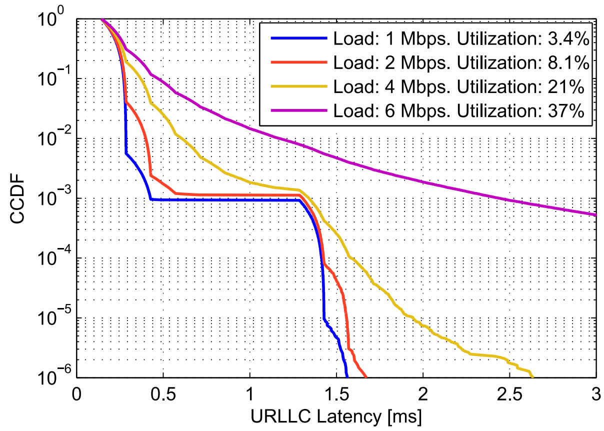

Fig. 3: Latency distribution at different offered loads. TTI length: 0.14 ms. HARQ RTT: 8 TTIs. BLER: 0.1%. $\alpha = 0.01$.

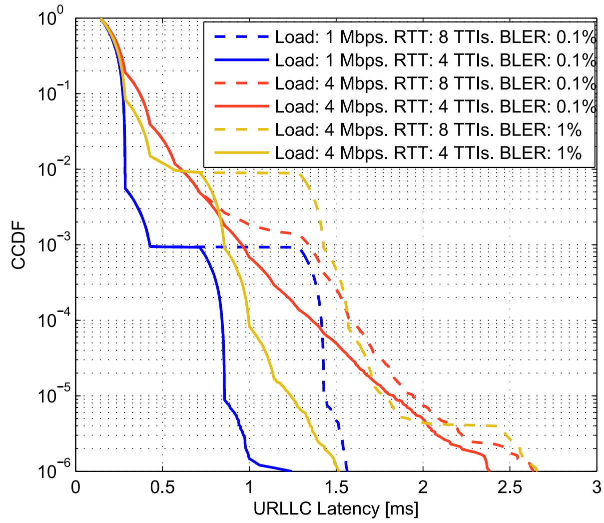

Fig. 4: Latency distribution with different configurations of the HARQ RTT and BLER target. TTI length: 0.14 ms. $\alpha = 0.01$.

the larger control overhead when scheduling with short TTIs, which results in a lower availability of radio resources for URLLC data transmissions. Note that the case with 0.5 ms TTI experience the lowest PRB utilization. Although the signalling overhead is larger as compared to the 1.0 ms TTI case, the 0.5 ms TTI have the advantage of higher resource granularity, which results in more resource-efficient scheduling of the small URLLC data payloads.

2) Interference filtering: Fig. 3 shows the latency distribution at different offered loads, and including the benefits of interference filtering (explained in Section III-C). Only cases with a TTI duration of 0.14 ms are presented, given its large latency-reduction potential. Comparing the 1 Mbps performance in Fig. 3 with the one shown in Fig. 2, lower resource utilization and a 1.9 ms latency reduction (from 3.3 ms to 1.4 ms) is obtained at the 10^{-5} percentile, which can be attributed to the more accurate link adaptation. Although not shown, the benefits of interference filtering become even greater at higher offered load, e.g. 2 or 4 Mbps per cell, when the cell activity is higher and more sporadic interference is experienced in the network. For loads up to 4 Mbps, good latency performance (below 2 ms) is obtained despite the relatively high system utilization. As we further increase the load to 6 Mbps, non-negligible queuing starts to occur at the cells' buffers, which deteriorates considerably the achievable latency.

3) HARQ RTT and BLER optimization: Fig. 4 show the performance with different configurations of the HARQ RTT and the BLER target, as discussed in Section III. At 1 Mbps load, the HARQ-related processing delay is the dominant component of the total latency. Hence, by reducing the HARQ RTT to 4 TTIs, the 1 ms latency and 99.999% reliability required for URLLC is fulfilled. The case with 4 Mbps offered load and 0.1% BLER target does not experience significant improvement when reducing the HARQ RTT, since the queuing delay is the dominant component. Instead, it is beneficial to operate at higher BLER target (1%) in order to increase the spectral efficiency of the system and reduce the queue length. It can be observed that, after reducing the experienced queuing delay, the benefit of short HARQ RTT becomes more evident.

4) Global performance summary: Fig. 5 shows the 10^{-5}-percentile latency performance at different loads when applying the proposed enhancements. The top of each bar indicates the achieved latency without any of the proposed enhancements, i.e. a fixed BLER target of 0.1% for all loads, $\alpha = 1$, and a HARQ RTT of 8 TTIs. The latency varies from 3.3 ms, at 1 Mbps offered load, to $\sim$100 ms for a offered load of 8 Mbps. The bottom of each bar indicates the optimized latency performance, and each segment represents the latency improvement provided by a certain enhancement. It is observed that the URLLC requirements are fulfilled for loads up to 2 Mbps (latency of 0.98 ms), but cases with 4 or 6 Mbps offered load also experience decent performance (1.24 ms and 1.63 ms, respectively). The relative gain of the proposed latency and reliability improvements depend on the system load. At low load, the processing time is one of

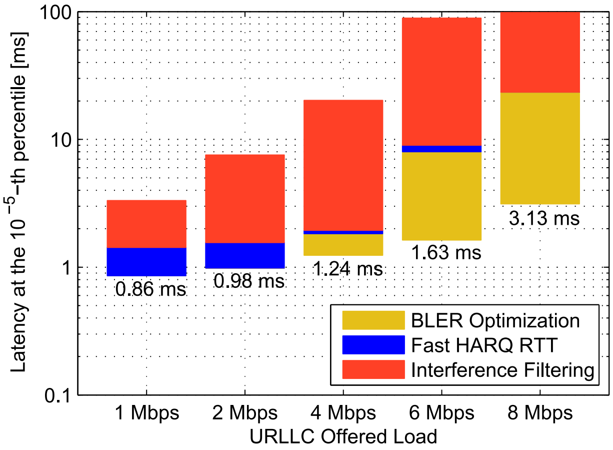

Fig. 5: Achievable latency at the 10^{-5} percentile under different offered loads. The performance improvement provided by the evaluated techniques is also presented. TTI length: 0.14 ms. $\alpha = 0.01$.

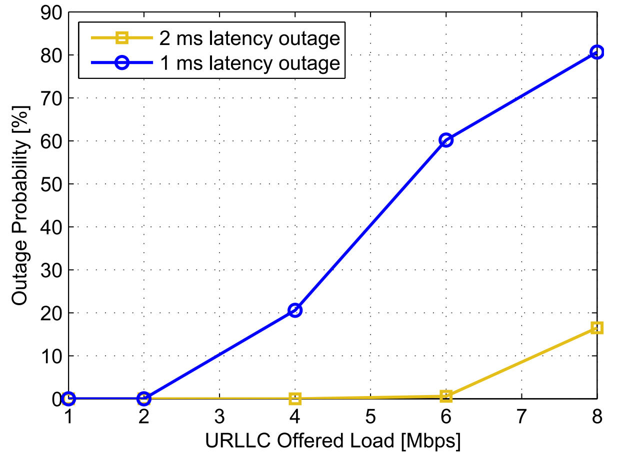

Fig. 6: Percentage of UEs not satisfying a certain latency requirement at the 10^{-4} percentile. TTI length: 0.14 ms. $\alpha = 0.01$.

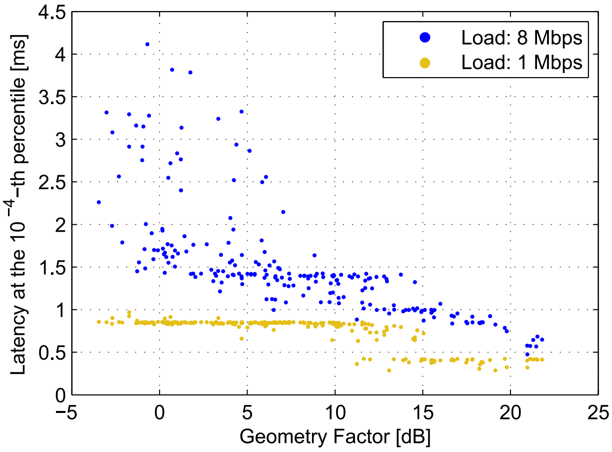

Fig. 7: Scatter plot of the per-user latency performance at the 10^{-4} percentile versus the Geometry Factor. TTI length: 0.14 ms. $\alpha = 0.01$.

the dominant components of the achievable latency; therefore, significant benefit is obtained by reducing the HARQ RTT. As we increase the load, the cell activity starts to increase resulting in rapidly-varying interference conditions. As a consequence, interference filtering provides large gains, as low-pass information of the experienced interference is implicitly included in the CQI report. On top of this, there is also relevant gain from using a higher BLER target at high load in order to reduce the queuing delay experienced in the cells (see Fig. 4). Specifically, the performance at 4, 6 and 8 Mbps offered load is improved by increasing the BLER target from 0.1% to 1%, 5% and 10%, respectively[2].

5) Per-User performance analysis: The results in Fig. 2-5 show the latency statistics from all the UEs in the network. However, due to the different coverage conditions of the UEs, it is likely to happen that not all the UEs experience the same performance. In order to quantify these effects, we analyse the latency performance for each UE, and determine the ratio of UEs which do not satisfy a certain latency requirement (i.e. outage probability). The latency is analysed at the 10^{-4} percentile, since the amount of samples available per UE is lower as compared to the global analyses. Fig. 6 shows the UE outage for different offered loads. The HARQ RTT is set to 4 TTIs, whereas the BLER target is configured for each load according to what provides the best latency performance. It is observed that a 1 ms latency with 99.99% reliability can be achieved by all the simulated UEs for loads up to 2 Mbps. As we increase the load, the outage probability drastically increases, e.g. 60% outage probability at 6 Mbps offered load. For a more relaxed latency constraint of 2 ms, the outage is significantly reduced, being no larger than 20% in any of the evaluated load conditions.

Intuitively, there is some correlation between the UE latency performance and its coverage conditions. Fig. 7 shows a scatter plot of the per-user latency performance at the 10^{-4}-percentile versus its Geometry factor (Γ). The Geometry factor is equivalent to the average signal to interference-and-noise ratio (SINR) experienced by the UE under full-load conditions; i.e. $\Gamma = P_i / (\sum_{n \neq i} P_n + N_0)$, where P_n is the received signal power from cell n, i denotes the serving cell, and N_0 is the thermal noise power. As expected, users located in cell-edge areas (low Γ) generally experience worse performance as compared to cell-center UEs. At 1 Mbps, the per-user latency performance can be grouped into two groups: (i) users with very good channel quality which do not experience any re-transmissions in the evaluated percentile, and therefore achieve a latency of only $\sim$0.4 ms; and (ii) users in less favourable channel conditions, experiencing typically one HARQ retransmission at the 10^{-4} level (resulting in a latency of $\sim$0.86 ms, as discussed in Section III-B). At 8 Mbps load, the achievable latency is largely affected by the queuing delay. The queuing

[2]The optimal BLER target for each load is obtained heuristically using simulations.

delay is not constant but naturally varies in accordance to the instantaneous variations of the incoming traffic. This results in a latency performance that varies from user to user but clearly correlated with the UE-specific experienced channel quality. The presented per-user performance statistics justify recent discussions in standardization proposing to perform admission control to discard UEs that are highly likely to not fulfil the required quality of service (due to the UE coverage conditions, system load, etc.) [10].

VI. Conclusions

Motivated by the traffic characteristics and stringent requirements of URLLC, we have presented different MAC layer enhancements for supporting URLLC in cellular networks. It is shown that the link adaptation imperfections, as a consequence of the very sporadic traffic, can be reduced by including low-pass filtered interference information in the CQI report, whereas short TTI and faster processing at the UE and base station is of significant importance to reduce the delay during HARQ retransmissions. Extensive system-level simulations have been carried out in order to evaluate the benefit of the proposed solutions. It has been shown how latencies below 1 ms with the required 99.999% reliability are achieved at low load scenarios, whereas some performance degradation (1 - 3 ms latency) is experienced at higher loads as a consequence of the higher queuing delay and inter-cell interference. The latency performance has also been evaluated on a per-user basis. The percentage of users not satisfying the 1 ms latency requirement drastically increases with the load, e.g. 20% and 60% for 4 and 6 Mbps offered load, respectively. In this regard, strong correlation is observed between the user latency performance and its experienced channel quality.

Our current work focuses on further enhancements to the link adaptation and scheduling mechanisms, as well as cases with mixed traffic classes, e.g. URLLC and MBB.

Acknowledgment

Part of this work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660) 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 TR 38.913 v14.1.0, "Study on scenarios and requirements for next generation access technologies", January 2017.

[2] P. Popovski, "Ultra-reliable communication in 5G wireless systems", *International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014.

[3] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, "Signal quality outage analysis for ultra-reliable communications in cellular networks", *IEEE Globecom Workshops*, Dec. 2015.

[4] F. Kirsten, D. Ohmann, M. Simsek and G. P. Fettweis, "On the utility of macro- and microdiversity for achieving high availability in wireless networks", *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept. 2015.

[5] N. Brahmi, et al., "Deployment strategies for ultra-reliable and low-latency communication in factory automation", *IEEE Globecom Workshops*, Dec. 2015.

[6] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson and J. Ansari, "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G", *IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sept. 2016.

[7] 3GPP TR 36.881 v14.0.0, "Study on latency reduction techniques for LTE", Jun. 2016.

[8] G. Pocovi, K. I. Pedersen, B. Soret, Mads Lauridsen and P. Mogensen, "On the impact of multi-user traffic dynamics on low latency communications", *International Symposium on Wireless Communication Systems (ISWCS)*, Sept. 2016.

[9] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji and R. Jäntti, "Link adaptation design for ultra-reliable communications", *IEEE International Conference on Communications*, May 2016.

[10] R1-1609545, "Remaining details of URLLC system level evaluation assumptions", 3GPP TSG-RAN WG1 #86bis, Oct. 2016.

[11] 3GPP TR 38.802 v2.0.0, "Study on new radio access technology physical layer aspects", March 2017.

[12] K. I. Pedersen, G. Berardinelli, F. Frederiksen and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases", *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53-59, March 2016.

[13] R1-1609664, "Comparison of slot and mini-slot based approaches for URLLC", 3GPP TSG-RAN WG1 #86bis, Oct. 2016.

[14] D. Laselva et al., "On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", *IEEE Vehicular Technology Conference*, Sept. 2009.

[15] H. Holma and A. Toskala, "LTE Advanced: 3GPP Solution for IMT-Advanced", John Wiley & Sons Ltd, 2011.

[16] G. Berardinelli, S. Khosravirad, K. I. Pedersen, F. Frederiksen and P. Mogensen, "Enabling early HARQ feedback in 5G networks", *IEEE Vehicular Technology Conference*, May 2016.

[17] A. Duran, M. Toril, F. Ruiz and A. Mendo, "Self-Optimization algorithm for outer loop link adaptation in LTE", *IEEE Communications Letters*, vol. 19, no. 11, pp. 2005-2008, Nov. 2015.