IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 18, 2022, accepted March 4, 2022, date of publication March 10, 2022, date of current version March 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158361

On the Cost of Achieving Downlink Ultra-Reliable Low-Latency Communications in 5G Networks

GUILLERMO POCOVI, TROELS KOLDING, AND KLAUS I. PEDERSEN, (Senior Member, IEEE)
Nokia Standards, 9220 Aalborg, Denmark
Corresponding author: Guillermo Pocovi (guillermo.pocovi@nokia-bell-labs.com)

ABSTRACT We determine the cost of serving Ultra-Reliable Low-Latency Communications (URLLC) traffic (1 ms one-way latency, 99.999% reliability) in 5G New Radio (NR) Macro cellular networks. The cost is measured as the degradation of the enhanced Mobile Broadband (eMBB) downlink system capacity when serving a certain offered load of URLLC traffic on the same radio carrier. A methodology for assessing the cost of URLLC is presented, which takes into account all the aspects related to configuring a suitable resource allocation and link adaptation strategy, as well as the associated control channel overhead due to the use of mini-slots with very frequent control channel resources for monitoring downlink data assignments. When taking a holistic view on the performance, advanced system-level simulation results show that 1 Mbps of URLLC traffic results in an eMBB throughput reduction of up to 60 Mbps, i.e. URLLC traffic can be up to 60 times more costly than traditional best-effort. The presented results can be used by cellular service providers, such as operators, for understanding and dimensioning the tradeoffs and impact towards traditional network services when adding URLLC services to their portfolio.

INDEX TERMS Ultra-reliable low-latency communications, 5G New Radio.

I. INTRODUCTION

Ultra-Reliable Low-Latency Communications is envisaged to provide unprecedented levels of latency and reliability for wireless transmission of data over private and public cellular networks. The 5G New Radio (NR) standard has been designed from scratch (i.e. since its initial release, Release 15 [1]), with the target of supporting one-way downlink (DL) or uplink (UL) user-plane latencies of 1 ms at a reliability of at least 99.999% as required for use-cases such as industrial automation, intelligent transportation, and remote healthcare. 5G NR Releases 16 and 17 added support for further reduced latency down to 0.5 ms and an increased reliability up to 99.9999% [2], [3]. As per [1], the reliability is defined as the probability that a data packet (typically of small size) is successfully received within the given latency requirement.

To provide URLLC, the NR standard supports a large set of features [1], [4], [5] such as a short transmission time interval (TTI), down to tens of μs, accelerated processing for the encoding and decoding of data at both the User Equipment (UE) and base station (gNB), improved UE channel quality indicator (CQI) feedback to allow link adaptation decisions targeting down to 10^{-5} block error rate (BLER) in the air interface, as well as proactive and reactive repetition or retransmission schemes, e.g. Hybrid Automatic Repeat Request (HARQ), to allow retransmitting the data without exceeding the available latency budget. Naturally, in many cases, such features provide improvements to the latency and/or reliability at the expense of a degradation of the spectral efficiency [6]. As an example, short TTIs are known for increasing the control overhead since scheduling grants are required more often [7]; similarly, targeting 10^{-5} BLER of the data transmissions imply using a lower modulation and coding scheme (MCS) with respect to what is generally used for enhanced Mobile Broadband (eMBB) services (with a typical BLER target in the order of 10%), hence having a negative effect on the spectral efficiency [8].

The majority of these URLLC enablers, while studied exhaustively in literature (e.g. analytically or via computer simulations), are not yet commonly deployed in real wide area networks as initial 5G NR deployments are mainly targeting traditional mobile broadband applications. Nevertheless, the commercial interest on URLLC is continuously increasing and it is expected that more and more

The associate editor coordinating the review of this manuscript and approving it for publication was Yu Wang.

URLLC-specific features be introduced in both UE and network side to gradually improve the achievable latency and reliability performance.

With this in mind, the goal of this article is to understand how expensive URLLC traffic is compared to traditional best-effort mobile broadband traffic, as well as to quantify the expected degradation of the eMBB capacity when introducing certain amount of URLLC traffic in the network. Coexistence of URLLC traffic and eMBB traffic on the same radio carrier has been analyzed in several studies [9]–[14]. For instance, the articles in [9], [10], [12], and [14] present different strategies for joint scheduling of URLLC and eMBB traffic on the same radio carrier. Majority of these assume the so-called punctured-scheduler supported by 5G NR specifications, where eMBB data is scheduled on a long TTI (e.g. of 0.5-1.0 ms duration), while URLLC urgent transmissions may be transmitted with a short TTI by overwriting part of the ongoing eMBB transmissions [9]. As reported in [9], puncturing of eMBB transmissions significantly reduces the probability of successfully decoding the corresponding DL data which results in a degradation of the eMBB latency and throughput. To deal with this, [10] proposes to favour puncturing of high signal to interference-and-noise ratio (SINR) eMBB users, in order to protect eMBB users with low data rate. Another approach, which is considered in this work, is to schedule both URLLC and eMBB transmissions with a short TTI duration (e.g. 143 μs or less) and multiplexed over different physical resource blocks (PRBs) using frequency-division multiplexing [13]. This approach avoids the need for puncturing eMBB data transmissions (and corresponding performance degradation) but it comes at the expense of larger control overhead since the gNB needs to issue scheduling grants for eMBB traffic more often. The paper in [15] uses queuing theory models to derive the URLLC capacity (admissible load subject to QoS constraints) with respect to various system parameters such as the link SINR, system bandwidth, and the packet latency and reliability requirements. Despite these numerous studies, none of these provide a comprehensive quantitative analysis on the degradation of the eMBB network capacity with respect to the URLLC load. Some performance numbers could be extrapolated from existing eMBB throughput and/or spectral efficiency analyses presented in e.g. [10], [11]; however, these studies do not accurately model all the aspects having an impact on the system and end-user performance, e.g. control channels and associated overhead.

Given this gap in the literature, the contributions of this article are the following:

- We provide key recommendations on the configuration of the Physical layer (PHY) and radio resource allocation strategies that need to be adopted for supporting URLLC. Particularly, the focus is on the transition from an *eMBB-centric* cellular network (as the majority of 5G deployments today) to a network capable of serving URLLC and eMBB traffic types simultaneously.
- We present an evaluation methodology and associated key performance indicators (KPIs) for determining the cost of providing URLLC-grade service in cellular networks. The presented KPIs take into account the implications of adopting a URLLC-suitable resource allocation strategy, including the effects of the control channel overhead which is found to play a major role in the performance.
- The presented evaluation methodology is applied to quantify the cost for a Urban Macro cellular network scenario. To this end, system-level simulations with high degree of realism are conducted, where we model a multi-cell, multi-user and mixed-service Macro network environment, including the effects of time-varying traffic and interference, link adaptation imperfections, and control and reference signal overhead. The simulation assumptions and methodology are aligned and calibrated with the 3GPP Release-16 NR evaluation guidelines outlined in [16] and [17].

The rest of the paper is organized as follows: First, in Section II, we present the considered network and UE deployment assumptions, including the adopted URLLC and eMBB traffic models. This section also describes the developed evaluation methodology and the KPIs of interest. Section III discusses the PHY and medium access control (MAC) layer configurations, resource allocation strategies and assumptions for meeting the URLLC requirements. Performance results are presented in Section IV, followed by Conclusions in Section V.

II. SETTING THE SCENE

A. NETWORK AND TRAFFIC MODEL

We consider a traditional urban macro (UMa) cellular network with $C = 21$ cells, deployed in a sectorized manner on a hexagonal grid; i.e. 7 sites with three cells each (covering a 120 degree sector) and 500-meter inter-site distance. A wrap around mechanism is adopted in order to provide statistically equivalent interference for all cells in the simulation area [18]. The wrap-around mechanism consists of placing an additional set of macro cells outside the simulation area (also positioned following the hexagonal grid) that generate interference with the same statistical properties (in both time and frequency) as the cells inside the simulation area. A set of $U_u = 10$ URLLC UEs and $U_e = 5$ eMBB UEs are deployed stationary in each cell according to a spatial uniform distribution. The URLLC traffic is modeled as small payloads of B_u= 32 or $B_u = 200$ [Bytes/packet], which are generated for each URLLC UE in the DL direction following a Poisson arrival process with a mean arrival rate λ_u [packet/s]. The offered load of URLLC traffic per cell is given by $L_u = U_u \cdot B_u \cdot \lambda_u \cdot 8 = B_u \cdot \lambda_{cell} \cdot 8$ [bps], where 8 corresponds to the conversion from Bytes to bits, and $\lambda_{cell} = U_u \cdot \lambda_u$ is the mean arrival rate of URLLC packets per cell. The eMBB traffic is modelled as full-buffer traffic with infinite payload size. That is, the downlink transmission buffers for each eMBB UE are assumed to contain an unlimited amount of data to be

delivered to the UE, e.g. resembling the download of a large file size [19]. The average data rate of the u_e-th eMBB UE served by cell c, $r_{u_e,c}$ [bps], is measured as the total amount of correctly-received bits $B_{u_e,c}$ divided by the simulation time T [s], i.e.:

$$r_{u_e,c} = \frac{B_{u_e,c}}{T} , \tag{1}$$

B. EVALUATION METHODOLOGY AND KPIs

Simulations are run for different offered loads L_u of URLLC traffic. For each offered load, we determine the average carried eMBB throughput per cell R_e, i.e. the sum of the average data rate experienced by each eMBB UE:

$$R_e = \frac{1}{C} \sum_{c=1}^{C} \sum_{u_e=1}^{U_e} r_{u_e,c} . \tag{2}$$

The cost is determined as the relative reduction of total carried eMBB throughput per cell R_e with respect to the URLLC offered load L_u, i.e.

$$\phi = \frac{R_e(L_{u,2}) - R(L_{u,1})}{L_{u,2} - L_{u,2}}. \tag{3}$$

For instance, the cost is said to be $\phi = -10$, if increasing the URLLC offered load in a cell from $L_{u,1} = 0$ Mbps $\rightarrow L_{u,2} =$ 1 Mbps, results in a reduction of 10 Mbps of the eMBB cell throughput R_e.

In addition to the cost, another KPI of interest is the one-way downlink latency experienced by each URLLC packet. The latency is measured from the moment a URLLC payload arrives at the serving cell until it is successfully received at the UE. Assuming that the URLLC payload can be entirely scheduled on a single TTI (which is a reasonable assumption for small payload sizes), the latency of a successfully-received URLLC packet τ_0 [seconds] equals

$$\tau_0 = t_q + t_{gNB} + t_a + t_{TTI} + t_{UE} , \tag{4}$$

where t_q is the queuing delay of the URLLC payload at the gNB; t_{gNB} and t_{UE} is the processing time at the gNB and UE, respectively; t_a is the so-called TTI alignment which is typically bounded in the interval $[0, t_{TTI}]$ for frequency-division duplexing (FDD) systems; and t_{TTI} is the over-the-air transmission time, essentially corresponding to the TTI duration.

For cases where the first transmission is erroneously decoded by the UE, a HARQ retransmission is triggered. Each retransmission adds additional τ_{HARQ} to the latency of the URLLC payload, where τ_{HARQ} denotes the HARQ round trip time (RTT). The URLLC latency experienced after a successful N-th retransmission equals,

$$\tau_N = \tau_0 + N \cdot \tau_{HARQ} . \tag{5}$$

Note that (5) assumes that HARQ retransmissions are always prioritized, and hence are not subject to queuing delays. In the following, the considerations for selecting t_{TTI}, among other physical layer (PHY) parameters, are provided.

III. PHYSICAL LAYER AND RESOURCE ALLOCATION CONSIDERATIONS

We follow the 5G NR PHY assumptions as outlined in [20], focusing primarily on the DL performance. Users are dynamically multiplexed on a time-frequency grid of resources using orthogonal frequency division multiple access (OFDMA) and FDD duplexing. We assume the PHY setting with 30 kHz subcarrier spacing (SCS) as commonly used in today's wide area networks, resulting in a OFDM symbol (OS) duration $t_{OS} = 35.71\ \mu s$. The carrier bandwidth is 40 MHz deployed in the 4 GHz frequency band.

The 5G NR standard allows different TTI durations ranging from a *slot*, composed of 14 OFDM symbols; to *mini-slots* of 1 to 13 OFDM symbols [20]. For initial 5G NR deployments targeting mainly eMBB services, slot-based scheduling is generally used as illustrated in Fig. 1(A), where the physical downlink control channel (PDCCH) is located in the first 1-2 symbols in the slot, and the data resources, known as the physical downlink shared channel (PDSCH), are located in the remaining 12-13 symbols of the slot. For the assumed 30 kHz subcarrier spacing, the transmission duration corresponds to $t_{TTI} = 14 \cdot t_{OS} = 0.5$ ms. Downlink control information (DCI) is transmitted on the PDCCH which indicates to each individual user the location (in time and frequency) of the user-specific PDSCH resources, the MCS that is used, among other transmission parameters needed to decode the data. As shown in Fig. 1(A), other signals such as Demodulation Reference Signals (DMRS) may also be transmitted for the purpose of channel estimation for demodulation, and for generating channel state information (CSI) feedback.

However, when attempting to support even a single URLLC user with stringent requirements of 1 ms latency and 99.999% reliability, the resource allocation format and PHY signals structure need to change significantly. For instance, if keeping the resource allocation structure in Fig. 1(A), the sum of the t_{TTI} and (worst-case) t_a component (i.e. $t_a = t_{TTI}$) in (4) already consumes the 1 ms latency budget thus not leaving any room for gNB/UE processing times, potential queueing delay and/or HARQ retransmissions. To reduce the latency, not only shorter TTI durations need to be used, but also frequent PDCCH resources for monitoring downlink data assignments need to be provided to the UE. Based on the analysis in [21], Fig. 1(B) illustrates URLLC-suitable PHY configuration with 7 PDCCH resources evenly distributed in the 14 OS slot and a mini-slot duration $t_{TTI} = 2 \cdot t_{OS} = 71.4$ μs. Each mini-slot contains DMRS reference signals used for channel estimation as well as PDCCH resources which provide the scheduling information (DCI) to one or more UEs that are scheduled on separate PRBs (using OFDMA) in the mini-slot. Note that having PDCCH resources in each mini-slot is essential in order to keep the TTI-alignment component in (4) in the $t_a \in [0, t_{TTI}]$ interval. The presented slot structure leverages the flexible PHY framework of NR, where the gNB can configure one or multiple control channel resource sets (CORESET) flexibly on different parts of the

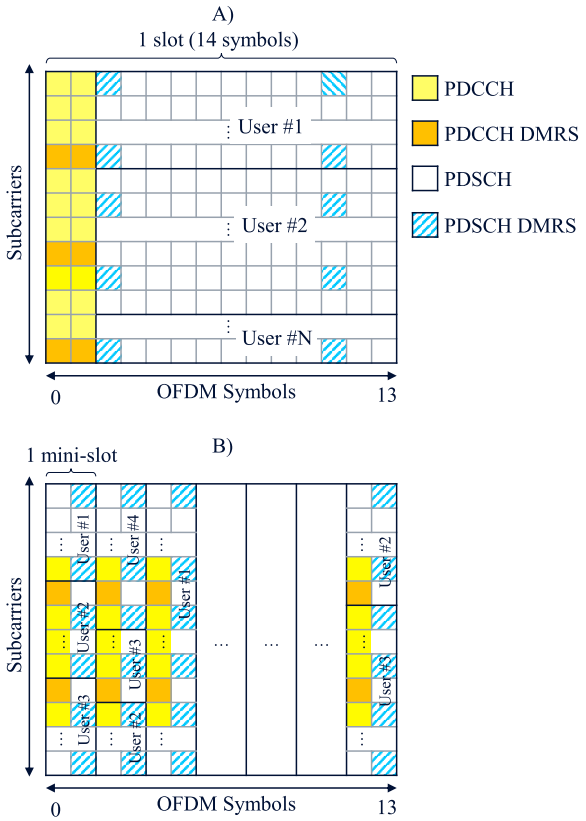


FIGURE 1. Default 5G NR frame structure (A) and optimized frame structure to support URLLC (B).

time-frequency resource grid. In Table 1, it is shown that this configuration allows to achieve the 1 ms latency requirement even after 1 HARQ retransmission, while still leaving some room for queuing delay t_q and potential fragmentation of the payload transmission across multiple TTIs. We refer to [21] for additional latency budget calculations with different mini-slot durations and SCS configurations.

A. CONTROL CHANNEL OVERHEAD CONSIDERATIONS

With the above in mind, a short TTI duration of 2 OFDM symbols is adopted for scheduling both URLLC and eMBB services.[1] The PDCCH is located on every second symbol of the radio slot as depicted in 1(B). In the frequency domain, the bandwidth of each PDCCH resource needs to be sufficiently large to accommodate the DCI scheduling the transmission of the URLLC payload in line with the number of URLLC payloads expected to be served during each 2 OS scheduling interval. Table 2 shows the amount of overhead resources assumed with respect to the URLLC arrival rate in each cell λ_{cell}. For reference, the case with no URLLC traffic and full-slot scheduling (Fig. 1(A)) is also included. For such case, a PDCCH occupying 50% of 1 OS is assumed to allow

[1]Note that another alternative is to apply the so-called punctured scheduling scheme, where eMBB data is scheduled on a long 14 OS TTI, while URLLC urgent transmissions are transmitted with a short TTI by overwriting part of the ongoing eMBB transmissions [9]. However, as reported in [22], for medium to high loads of URLLC traffic, the approach considered in this work is preferable in terms of the eMBB performance.

TABLE 1. URLLC latency budget excluding the effects of queuing delay. Processing and HARQ retransmission times are as per [21].

Component	Description	Value [ms]
t_{gNB}	BS Tx processing delay	0.098
t_a	Frame alignment (worst-case)	0.071
t_{TTI}	TTI for data packet transmission	0.071
τ_{HARQ}	a) UE Rx processing delay and HARQ feedback preparation	0.196
	b) Alignment to control opportunity	0.018
	c) Transmission of the HARQ-ACK	0.036
	d) BS processing delay	0.196
	e) Frame alignment	0.018
	f) TTI for data packet transmission	0.071
t_{UE}	UE Rx processing delay	0.116
τ_0	Latency without retransmission ($t_{gNB} + t_a + t_{TTI} + t_{UE}$)	0.357
τ_1	Latency with 1 retransmission ($\tau_0 + \tau_{HARQ}$)	0.893

scheduling of up to 7 frequency-multiplexed eMBB UEs in the slot (more details on the PDCCH overhead calculation are explained next). When using mini-slot scheduling, the number of schedulable eMBB users per 2 OS mini-slot is reduced to 1 in order to keep the PDCCH overhead low; this essentially allows up to 7 eMBB UEs to be scheduled in the same slot using time-domain multiplexing instead of frequency-domain multiplexing. For the cases with URLLC traffic, the PDCCH overhead varies between 1.75 and 4.375 (out of 14 OS) per slot depending on the actual URLLC offered load. In addition to the PDCCH, PDSCH DMRS reference signals are assumed to occupy 50% of the second OFDM symbol in each mini-slot (see Fig. 1(B)) resulting in a total control overhead of up to 56.25% [23]. The higher the overhead, the less resources available for the PDSCH (data) channel thus having an impact on the system performance.

The PDCCH overhead calculations in Table 2 are based on allocating sufficient PDCCH resources such that, with 99.999% probability, all URLLC users with data to transmit (i.e. M_u URLLC UEs) can be scheduled on the next available 2 OS TTI. In each mini-slot, one or more eMBB UEs (denoted as M_e) are also scheduled only if less than M_u URLLC UEs are served in the mini-slot. The value of M_u is determined using the well-known mathematical properties of a Poisson arrival process [24], where the the probability P_m of m packet arrivals in a time interval ΔT is given as,

$$P_m = \frac{(\lambda_{cell} \cdot \Delta T)^m}{m!} \cdot \exp\left(-\lambda_{cell} \cdot \Delta T\right). \tag{6}$$

From (6), with a 99.999% probability, the maximum number of URLLC packet arrivals M_u expected over a $\Delta T = t_{TTI}$ scheduling interval can be determined as,

$$M_u = \arg\min_m \sum_{i=0}^{m} P_i \geq 0.99999\,. \tag{7}$$

TABLE 2. PDCCH and DMRS overhead assumptions with respect to the URLLC packet arrival rate per cell $\lambda_{cell} = U_u \cdot \lambda_u$.

λ_{cell} [packet/s]	Max # of schedulable eMBB UEs (M_e)	Max # of schedulable URLLC UEs (M_u)	Overhead per slot [# of OFDM symbols] PDCCH	DMRS	Total overhead
0	7 (full-slot)	0	0.5	2·0.5	1.5 (10.7%)
0	1 (mini-slot)	0	1.0	7·0.5	4.5 (32%)
500	$M_e + M_u = 2$		1.75	7·0.5	5.25 (37.5%)
1000	$M_e + M_u = 3$		2.625	7·0.5	6.125 (43.75%)
2000	$M_e + M_u = 4$		3.5	7·0.5	7.0 (50%)
4000	$M_e + M_u = 5$		4.375	7·0.5	7.875 (56.25%)

In accordance with the NR specifications, each DCI is transmitted on a number of control-channel elements (CCE) in the range {1, 2, 4, 8, 16}, where each CCE consists of 6 physical resource blocks (PRBs) in frequency and 1 OS in time, i.e. corresponding to 72 resource elements (REs). PDCCH overhead calculations in Table 2 are done assuming that each downlink data assignment (DCI), on average, occupies 1 CCE (72 REs) and 2 CCEs (2·72 REs) for eMBB and URLLC, respectively. The number of CCEs (also know as the *PDCCH aggregation level*) have been selected in accordance with the typical channel quality (SINR) experienced by the UEs in the simulations with respect to the NR link-level PDCCH performance reported in [25] and [26], and considering that URLLC UEs require much higher PDCCH decoding probability (≥99.9%), than what is required for the eMBB users (∼99%) [27].

Note that the presented overhead calculations are specific to the adopted NR numerology with 30 kHz SCS, where the use of mini-slot is mandatory to meet the stringent URLLC latency requirements. For higher SCS, e.g. 120 kHz or more, the support of URLLC may not result in a large increase in overhead (as compared to only serving eMBB traffic) since full-slot (14 OS) scheduling durations may be used for both URLLC and eMBB traffics (due to the natural latency reduction from shorter OFDM symbol duration and faster UE processing times [5]); However, please note that the use 120 kHz SCS or higher is typically reserved for Frequency Range 2 deployments (24 GHz carrier frequency or above) with typically small cell sizes which are not in the scope of this article.

IV. PERFORMANCE EVALUATION

A. SIMULATION ASSUMPTIONS

The evaluation is conducted via extensive dynamic system-level simulations following the 5G NR simulation methodology outlined in [16] and [17]. The simulation assumptions are summarized in Table 3. The network layout, UE distribution and traffic follow the description presented in Section II-A. The in-house developed system-level simulator has a time resolution of one OFDM symbol, and it includes explicit modelling of the majority of radio resource management functionalities such as dynamic packet scheduling and HARQ, as well as time- and frequency-varying inter-cell interference. The simulator has been used to generate a large variety of LTE and 5G NR performance results and has been calibrated with system-level simulators from several 3GPP

TABLE 3. Simulation assumptions.

Parameter	Value
Network env.	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance [17]
PHY settings	30 kHz subcarrier spacing; 12 subcarriers per PRB; TTI size of 2 OFDM symbols (71.4 μs)
Carrier config.	40 MHz carrier bandwidth (100 PRBs) at 4 GHz
Duplexing	Frequency division duplexing (FDD)
Control channel	Error-free PDCCH with load-dependent overhead as per Table 2
CQI/CSI configuration	CQI and PMI, reported every 2 ms with 2 ms processing delay; Sub-band size: 4 PRBs;
Antenna config.	4 x 4 single-user MIMO and MMSE-IRC receiver.
Packet scheduler	Proportional Fair; strict priority for URLLC traffic
HARQ	Async. HARQ with Incremental Redundancy; Max. 4 HARQ retransmissions. Processing time as in [21]
RLC	RLC Unacknowledged mode
UE distribution	10 URLLC UEs and 5 eMBB UEs per cell; UEs uniformly distributed in outdoor locations
Traffic model	URLLC: Poisson arrival of DL packets of size $B_u = 50$ and 200 B; Variable offered load per cell. eMBB: Full-buffer DL traffic

member companies, see e.g. [28] and [29]. On every TTI, the experienced SINR for each scheduled user is calculated per RE, assuming a minimum mean square error interference rejection combining (MMSE-IRC) receiver [30] and closed-loop 4 × 4 single-user MIMO. Given the SINR per RE, the effective exponential SINR model [31] is applied for link-to-system-level mapping to determine if the transmission was successfully decoded. Asynchronous adaptive HARQ with Incremental Redundancy is applied in case of failed transmissions.

Service-aware scheduling and link adaptation is considered in order to meet the stringent URLLC latency and reliability requirements, while maintaining as high spectral-efficiency as possible for the eMBB users [13]. That is, URLLC users are scheduled with a single Multiple-Input Multiple-Output (MIMO) spatial stream, i.e. benefiting from both transmission and reception diversity against fast fading and radio channel fluctuations, whereas dynamic rank adaptation is assumed for eMBB users allowing multiplexing of up to two spatial streams for favourable SINR conditions. Dynamic link adaptation is applied for the data transmissions on the PDSCH, based on periodical CQI reports from the UEs and using outer-loop link adaptation (OLLA) to reach a BLER target of 10^{-4} and 10^{-1} for URLLC and

eMBB UEs, respectively [32]. Also, the packet scheduler prioritizes URLLC transmissions and HARQ retransmissions over retransmissions and first transmissions of eMBB traffic. The maximum number of HARQ retransmissions is limited to 4 for both URLLC and eMBB UEs; although, in practice, URLLC transmissions experience at most one HARQ retransmission due to the low initial BLER target. The PDCCH containing the scheduling assignments is assumed to be error-free; however, the control channel overhead is adjusted in accordance with the offered load and the required PDCCH decoding probability for URLLC and eMBB users as described in Section III-A.

For each URLLC payload, the latency is measured from the moment it arrives at the serving cell until it is successfully received at the UE, as per the models described in Section II-B. For each simulated URLLC UE, the latency of each received URLLC payload is collected and the 99.999%-ile of the UE's experienced latency is determined. Ten different drops with randomized UE positions (i.e. $10 \cdot C \cdot U_u = 2100$ URLLC UEs in total) are simulated to ensure sampling of different coverage conditions. The simulation time corresponds to at least 1.000.000 successfully received URLLC payloads per simulated URLLC UE in order to ensure a reasonable confidence level for the considered performance metric [33].

B. PERFORMANCE RESULTS

We start by showing the latency and reliability performance achieved by URLLC UEs. Fig. 2 shows a scatter plot of the 99.999%-ile latency achieved by each simulated URLLC UE with respect to its pathgain towards its serving cell, for a fixed UE arrival rate of $\lambda_u = 50$ packet/s (i.e. $\lambda_{cell} = 500$ packet/s) and payload size of $B_u = 32$ B and $B_u = 200$ B. Many of the URLLC UEs, especially those experiencing good channel quality (low pathloss) towards the serving cell, achieve the 99.999% reliability target with a single transmission opportunity. This results in a latency of approximately 0.357 ms which is aligned with the latency breakdown in Table 1. For users with large pathloss to the serving cell, the achieved latency is generally worse as many of these users experience one or more HARQ retransmissions as well as fragmentation of the payload transmission across multiple TTIs which increases the 99.999%-level latency to 0.89 ms or more. It is also observed that the performance with $B_u = 200$ B payload is generally worse for cell-edge UEs, as they have more difficulty in keeping up with the generated data rate $B_u \cdot \lambda_u$ with the required latency and reliability.

In Fig. 3 we show the percentage of satisfied URLLC UEs for different offered loads of URLLC traffic L_u. As expected, the higher the offered load, the lower the amount of users meeting the requirements, mainly as a consequence of queuing delay and fragmentation of the payload transmission at the gNB. Interestingly, the case with $B_u = 200$ B payload offers better performance for the same offered load conditions (e.g. $L_u = 1$ Mbps); this is because, for the same offered load conditions, cases with $B_u = 200$ B have a lower arrival

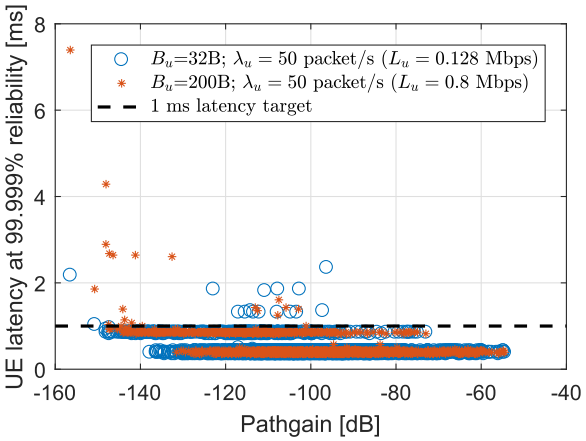


FIGURE 2. 99.999% of the achieved UE latency vs the UE's pathgain towards the serving cell for low offered load ($\lambda_u = 50$ packet/s) of URLLC traffic.

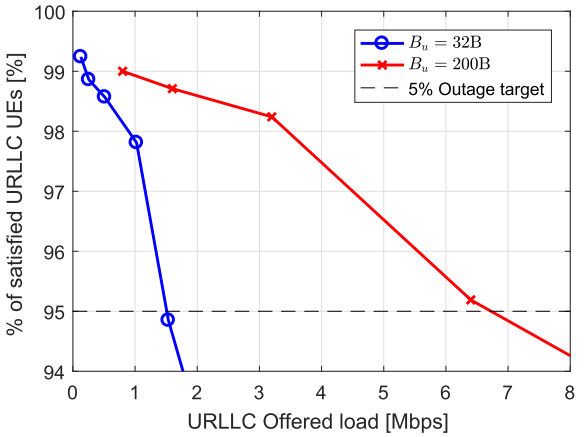


FIGURE 3. Percentage of URLLC UEs meeting the 1 ms and 99.999% latency and reliability requirements vs the URLLC offered load.

rate of packets which in turn result in lower control channel overhead. Also, $B_u = 200$ has the benefit from more robust coding due to larger codeblock size [34], and larger frequency diversity due to a larger PRB allocation which essentially improves the robustness against frequency-selective fading of the radio channel and interference. For an UE outage probability of 5% as defined in 3GPP in [16], approximately 1 and 6 Mbps of URLLC offered load can be supported for $B_u = 32$ B and $B_u = 200$ B payload size, respectively.

With respect to the impact towards traditional best-effort eMBB traffic, Fig. 4 shows aggregated eMBB cell throughput R_e with respect to L_u for both payload sizes of the URLLC traffic. The URLLC offered loads are limited to those where the UE outage probability of 5% is not exceeded as per Fig. 3. For each case, a minimum mean square error (MMSE) linear fit $y = \phi \cdot x + b$ is illustrated where the slope ϕ represents the relative reduction of eMBB throughput for each 1 bps of URLLC traffic. For small payload size of $B_u = 32$ B, the cost is close to -60, i.e. 1 Mbps of URLLC traffic reduces the eMBB cell throughput from ~150 to ~90 Mbps. Whereas the cost is significantly lower (approx. -11) for $B_u = 200$ B. This is mainly because larger payloads can be more efficiently transmitted than small payloads since the

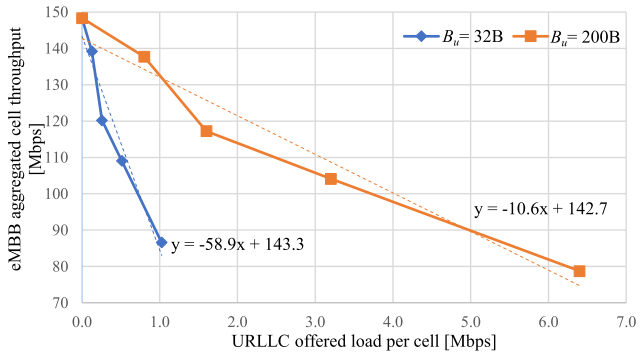


FIGURE 4. eMBB cell throughput vs URLLC offered load. Slope of the linear fit indicates the approximate cost ϕ of URLLC traffic as per (3).

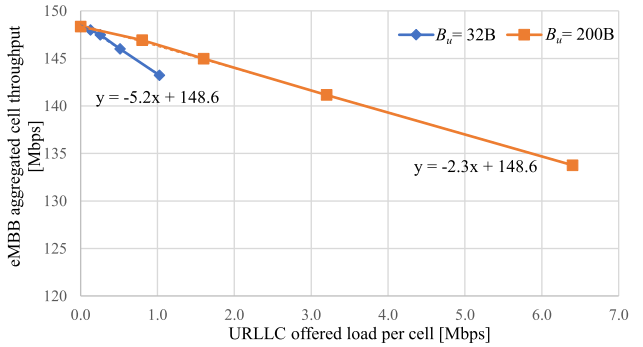


FIGURE 5. eMBB cell throughput vs URLLC offered load assuming fixed control channel overhead of 32%. Slope of the linear fit indicates the approximate cost from PDSCH (data channel) point of view.

amount of control channel resources occupied by each DL data assignment is the same regardless of whether a small or large amount of user data is scheduled.

It is worth highlighting that the cost has been so far quantified using the case with mini-slot scheduling and no URLLC traffic as the reference. In Table 4, we show the eMBB cell throughput statistics including also the case with full-slot scheduling illustrated in Fig. 1(A). In the absence of URLLC traffic ($\lambda_{cell} = 0$), it is shown that the switching from a traditional eMBB-suitable PHY configuration to a URLLC-optimized one already results in a significant ∼30% reduction (from 208 Mbps to 148 Mbps) of the cell capacity; this is a consequence of the larger overhead due to the frequent PDCCH and DMRS signals as described in Table 2. Serving some URLLC traffic in the system further degrades the performance; however, it is worth noting that the degradation of the eMBB throughput is not linear with respect to the URLLC offered load: for instance, for $B_u = 32$B, increasing L_u from 0 to 0.5 Mbps reduces the R_e by ∼40Mbps, whereas increasing L_u from 0.5 to 1 Mbps further reduces R_e by only ∼22 Mbps. The reason for this non-linear behaviour is the so-called *trunking efficiency* gain where control-channel resources can be more-efficiently dimensioned and utilized for high offered loads of URLLC traffic.

Finally, to better understand the contribution of the control (PDCCH) and data (PDSCH) channel on the overall cost, we show in Fig. 5 the aggregated eMBB cell throughput R_e with respect to L_u, but assuming the total overhead to be fixed to 32%. By fixing the control channel and reference signal overhead irrespective of the URLLC offered load, the observed slope indicates the cost from pure PDSCH (data channel) perspective, i.e. how much more data resources URLLC transmissions consume due to lower BLER target or more conservative link adaptation. It is observed that the cost is only −5.2 and −2.3 for the $B_u = 32$ B and $B_u = 200$ B payload size, respectively, as compared to −59 and −11 obtained in Fig. 4. Based on these observations, it is concluded that the control channel is one of the main culprits of the high cost of URLLC.

TABLE 4. eMBB cell throughput R_e vs URLLC offered load L_u.

λ_{cell} [packet/s]	$B_u = 32$ Byte		$B_u = 200$ Byte	
	L_u [Mbps]	R_e [Mbps]	L_u [Mbps]	R_e [Mbps]
0 (full-slot)	0.000	208.5	0.0	208.5
0 (mini-slot)	0.000	148.4	0.0	148.4
500	0.128	139.2	0.8	137.7
1000	0.256	120.2	1.6	117.2
2000	0.512	109.1	3.2	104.1
4000	1.024	86.7	6.4	79.0

V. CONCLUSION

In this paper we have evaluated the cost of providing URLLC-grade service (1 ms one-way latency, 99.999% reliability) in Urban Macro cellular networks. The cost has been measured in terms of the degradation of eMBB system capacity when serving a certain offered load of URLLC traffic on the same radio carrier.

When only taking into account the impact of the scheduling prioritization and more conservative link adaptation (i.e. excluding the impact of the control overhead), it is shown that 1 Mbps of URLLC traffic results in an eMBB cell throughput reduction of 2 to 5 Mbps, i.e. URLLC traffic is 2 to 5× times more costly than eMBB depending on the associated payload size of the URLLC traffic.

However, when accounting also for the larger control overhead due to the need for frequent and reliable PDCCH resources, the cost increases significantly to up to 60 times. In other words, the control channel overhead has a higher contribution than the data channel (PDSCH) on the quantified degradation of the system capacity. As an example, for a 40 MHz carrier bandwidth and 0 kbps of URLLC traffic, the eMBB carried throughput of the cell decreases from approximately 208 Mbps down to 148 Mbps (30% reduction) when switching from a 14 OFDM symbol (OS) scheduling interval to a 2 OS (i.e. mini-slot) scheduling interval. The eMBB cell throughput is further reduced to 139 Mbps (33% total reduction) when serving a URLLC offered load of 128 kbps in the downlink direction. The results show that URLLC load needs to be carefully tuned to how much eMBB capacity needs to be carried in a given cell.

Future work should also consider the cost for more relaxed service requirements, e.g. time-critical services with latency

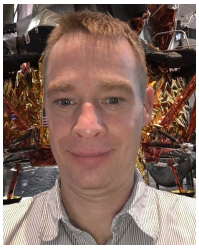
in the 2-5 ms interval and reliability in the order of 99.9% to 99.999%, as well as evaluating the cost in the UL direction.

REFERENCES

[1] *5G; Study on Scenarios and Requirements for Next Generation Access Technologies (Release 16)*, Standard TR 38.913, V16.0.0, 3GPP, Jul. 2020.
[2] S. Baek, D. Kim, M. Tesanovic, and A. Agiwal, "3GPP new radio release 16: Evolution of 5G for industrial Internet of Things," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 41–47, Jan. 2021.
[3] T.-K. Le, U. Salim, and F. Kaltenberger, "An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17," *IEEE Access*, vol. 9, pp. 433–444, 2021.
[4] S. Ye, "Support of ultra-reliable and low-latency communications (URLLC) in NR," in *5G and Beyond*, J. Fagerberg, D. C. Mowery, and R. R. Nelson, Eds. Cham, Switzerland: Springer, 2021, ch. 13, pp. 373–400.
[5] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Netw.*, vol. 32, no. 2, pp. 24–31, Mar. 2018.
[6] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.
[7] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.
[8] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 1005–1010.
[9] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–6.
[10] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "EMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.
[11] S. R. Pandey, M. Alsenwi, Y. K. Tun, and C. S. Hong, "A downlink resource scheduling strategy for URLLC traffic," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2019, pp. 1–6.
[12] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.
[13] 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. 28912–28922, 2018.
[14] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.
[15] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2411–2421, Nov. 2018.
[16] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, Standard TR 38.824 V16.0.0, 3GPP, Mar. 2019.
[17] *Study on Channel Model for Frequencies From 0.5 to 100 GHz (Release 16)*, Standard 38.901 V16.1.0, 3GPP, Tech. Rep., Nov. 2020.
[18] R. S. Panwar and K. M. Sivalingam, "Implementation of wrap around mechanism for system level simulation of LTE cellular networks in NS3," in *Proc. IEEE 18th Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2017, pp. 1–9.
[19] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2nd Quart., 2020.
[20] *NR and NG-RAN Overall Description; Stage-2*, Standard TR 38.300 V16.6.0, 3GPP, Sep. 2019.
[21] *Discussion on the RAN2 LS on TSN Requirements Evaluation*, Standard R1-1813120, 3GPP, Nov. 2018.
[22] G. Pocovi, "Radio resource management for ultra-reliable low-latency communications in 5G," Ph.D. dissertation, Dept. Electron. Syst., Aalborg Univ., Aalborg, Denmark, 2017.
[23] M. Enescu, *5G New Radio: A Beam-based Air Interface*. Hoboken, NJ, USA: Wiley, 2020.
[24] L. Kleinrock, *Theory, Queueing Systems*, vol. 1. Hoboken, NJ, USA: Wiley, 1975.
[25] V. Braun, K. Schober, and E. Tiirola, "5G NR physical downlink control channel: Design, performance and enhancements," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.
[26] H. Chen, D. Mi, M. Fuentes, E. Garro, J. L. Carcel, B. Mouhouche, P. Xiao, and R. Tafazolli, "On the performance of PDCCH in LTE and 5G new radio," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Jan. 2018, pp. 1–6.
[27] *Analysis of URLLC Reliability in DL HARQ*, Standard R1-1700265, 3GPP, Jan. 2017.
[28] *Study on Self Evaluation Towards IMT-2020 Submission*, Standard TR 37.910 V16.1.0, 3GPP, Nov. 2020.
[29] *Report on Evaluations for 5G-ACIA*, document RP-210490, Mar. 2021.
[30] M. Lampinen, F. D. Carpio, T. Kuosmanen, T. Koivisto, and M. Enescu, "System-level modeling and evaluation of interference suppression receivers in LTE system," in *Proc. IEEE 75th Veh. Technol. Conf. (VTC Spring)*, May 2012, pp. 1–5.
[31] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G. A. Seraji, "Link performance models for system level simulations of broadband radio access systems," in *Proc. IEEE 16th Int. Symp. Pers., Indoor Mobile Radio Commun.*, vol. 4, Sep. 2005, pp. 2306–2311.
[32] K. I. Pedersen, G. Monghal, I. Z. Kovacs, T. E. Kolding, A. Pokhariyal, F. Frederiksen, and P. Mogensen, "Frequency domain scheduling for OFDMA with limited and noisy channel feedback," in *Proc. IEEE 66th Veh. Technol. Conf.*, Sep. 2007, pp. 1792–1796.
[33] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen, and P. Mogensen, "Signal quality outage analysis for ultra-reliable communications in cellular networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.
[34] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

GUILLERMO POCOVI received the M.Sc. degree in telecommunications engineering from the Universitat Politècnica de Catalunya, in 2014, and the Ph.D. degree from Aalborg University, Denmark, in 2017. He is currently with Nokia Standards, Aalborg. His research activities include the standardization of ultra-reliable and low-latency communications (URLLC) and Industrial Internet of Things (IIoT) use cases in 5G New Radio.

TROELS KOLDING received the M.Sc. and Ph.D. degrees from Aalborg University, Denmark, in 1996 and 2000, respectively. His M.Sc. was achieved in collaboration with the Wireless Information Network Laboratory (WINLAB), NJ, USA. Since joining Nokia, in 2001, he has been active in research and management for standardization, network architecture, and portfolio management. He holds more than 50 granted U.S. patents and is an author of more than 80 scientific publications. His current research interests include 5G IIoT, time-sensitive communications, time-synchronization, and 5G/6G radio resource management and spectrum sharing.



KLAUS I. PEDERSEN (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. He is a Bell Labs Fellow at Nokia, currently leading the Radio Access Research Team in Aalborg, and a part-time External Professor at Aalborg University. He has authored publications on a wide range of topics, as well as an inventor on several patents. His current research interests include access protocols and radio resource management enhancements for 5G New Radio and its evolution to 5G-Advanced.