# Joint Link Adaptation and Scheduling for 5G Ultra-Reliable Low-Latency Communications

Pocovi, Guillermo; Pedersen, Klaus I.; Mogensen, Preben

*Published in:*
IEEE Access

*DOI (link to publication from Publisher):*
10.1109/ACCESS.2018.2838585

*Publication date:*
2018

*Document Version*
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 28, 2018, accepted May 4, 2018, date of publication May 21, 2018, date of current version June 19, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2838585*

# Joint Link Adaptation and Scheduling for 5G Ultra-Reliable Low-Latency Communications

**GUILLERMO POCOVI[1], (Member, IEEE), KLAUS I. PEDERSEN[1,2], (Senior Member, IEEE), AND PREBEN MOGENSEN[1,2], (Member, IEEE)**

[1]Nokia Bell Labs, 9220 Aalborg, Denmark
[2]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark

Corresponding author: Guillermo Pocovi (guillermo.pocovi@nokia-bell-labs.com)

This work was supported by the Innovation Fund Denmark under Grant 7039-00009B.

**ABSTRACT** This paper presents solutions for efficient multiplexing of ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) traffic on a shared channel. This scenario presents multiple challenges in terms of radio resource scheduling, link adaptation, and inter-cell interference, which are identified and addressed throughout this paper. We propose a joint link adaptation and resource allocation policy that dynamically adjusts the block error probability of URLLC small payload transmissions in accordance with the instantaneous experienced load per cell. Extensive system-level simulations of the downlink performance show promising gains of this technique, reducing the URLLC latency from 1.3 to 1 ms at the 99.999% percentile, with less than 10% degradation of the eMBB throughput performance as compared with conventional scheduling policies. Moreover, an exhaustive sensitivity analysis is conducted to determine the URLLC and eMBB performance under different offered loads, URLLC payload sizes, and link adaptation and scheduling strategies. The presented results give valuable insights on the maximum URLLC offered traffic load that can be tolerated while still satisfying the URLLC requirements, as well as what conditions are more appropriate for dynamic multiplexing of URLLC and eMBB traffic in the upcoming 5G systems.

**INDEX TERMS** 5G New Radio, link adaptation, scheduling, radio resource management, ultra-reliable low-latency communications.

## I. INTRODUCTION

The upcoming fifth generation (5G) New Radio (NR) will provide support for a wide range of services and applications [1], [2]. Besides enhanced mobile broadband (eMBB), supporting an evolution of today's broadband traffic, 5G will enable ultra-reliable low-latency communications (URLLC), where small payloads must be correctly transmitted and received in a very short time (up to 1 ms) with a success probability of 99.999% [3]. Support for such unprecedented requirements of latency and reliability will open the door to novel use cases, including wireless control and automation in industrial environments [4], inter-vehicular communication for safety [5], smart grids [6], and real-time tactile Internet services [7].

The strict requirements of URLLC call for a broad set of enhancements covering different parts of the 5G radio interface. The studies in [8]–[10] focus on addressing the harmful effects of the radio channel, which represent a major challenge to the reliability of the system. It is shown how a combination of micro- and macroscopic spatial diversity techniques plays an important role in dealing with the large- and small-scale fading effects, and the co-channel interference. To achieve low over-the-air transmission delay, the use of short transmission time intervals (TTIs) is of significant importance [11]. The study in [12] analyses the downlink latency performance with different TTI durations and load conditions where, in line with [13], the trade-offs between spectral efficiency, latency and reliability are observed. Under conditions of reliable control channels and feedback, the use of retransmission techniques such as hybrid automatic repeat request (HARQ) can substantially relax the block error probability (BLEP) constraint that the URLLC transmissions need to fulfil [14]. The advantage of using multiple transmission attempts, as compared to a single (very conservative) transmission, is a reduction of the average amount of radio resources required to transmit the small data

payloads [14]. Building on these studies, [15] presented multiple medium access control (MAC) layer enhancements for URLLC, which are corroborated by extensive system-level simulations of the downlink latency performance in a multi-user and multi-cell wide-area environment. In scenarios with only URLLC traffic, it is shown that the link adaptation inaccuracies, as a consequence of the very sporadic traffic and rapid interference variations, represent a major challenge.

The multiplexing of URLLC traffic with more traditional eMBB traffic is also gaining increased attention in industry and academia. In [16], a simplified analytical queuing analysis shows how dynamic multiplexing of both services (in both time and frequency domain) can significantly improve the resource efficiency of the wireless system. Following the same direction, [17] shows the benefits of punctured scheduling techniques, which allow the latency-critical traffic to overwrite the longer ongoing eMBB transmissions. Despite these valuable findings, there are still some challenges and open questions that require further study. Those include (i) how to efficiently distribute the resources between eMBB and URLLC while ensuring their respective quality-of-service (QoS) requirements, and (ii) how to deal with the larger inter-cell interference generated from scheduling eMBB users. The first challenge is typically addressed by the packet scheduler functionality. QoS-aware scheduling techniques for cellular systems such as High Speed Packet Access (HSPA) and Long Term Evolution (LTE) have been exhaustively studied in the open literature, see e.g. [18], [19]. It is clear from those studies that users with tight latency constraints should be prioritized when allocating radio resources, either by using hard priority or soft priority type-of solutions. Regarding (ii), co-channel inter-cell interference is one of the limiting factors in cellular networks, and has been addressed in many papers, e.g. [20], [21]. In the context of URLLC, [22] and [23] study inter-cell coordinated power boosting and/or cell muting schemes as a way to achieve high reliability and low latency in 5G, whereas [24] and [25] analyse different deployment strategies (cell layout and frequency reuse pattern) in order to meet the coverage requirements in a factory automation scenario. Reliable transmission of data under severe interference can also be handled to some extent by selecting a sufficiently robust modulation and coding scheme (MCS) such that low BLEP is achieved [26]. Selecting an appropriate BLEP target is, however, not trivial in dynamic multi-user environments, where multiple URLLC transmissions may need to be accommodated in the same TTI (with limited amount of frequency resources) [15].

Given the aforementioned challenges, this paper presents enhancements for efficient support of URLLC and eMBB in 5G cellular networks, focusing on the downlink. The proposed solutions are derived for a highly-dynamic environment, with multiple users and cells, time-varying traffic and inter-cell interference. Building on the previous work in [15], we consider the case where the network carries only URLLC traffic, and cases where URLLC user equipments (UEs) coexist with traditional best-effort eMBB traffic. As it will be shown, the challenges for achieving the stringent requirements of URLLC are rather different for those two cases, calling for different solutions.

The contributions of this paper can be summarized as follows:

- We present a QoS-aware and radio-channel-aware packet scheduling mechanism, able to efficiently serve the URLLC users in accordance with their QoS requirements, even in the presence of eMBB traffic.
  Our packet scheduling framework closely interacts with the link adaptation functionality, such that the BLEP of the URLLC transmissions is dynamically adjusted in coherence with the instantaneous URLLC load experienced per cell.
- We propose an attractive channel quality indicator (CQI) measuring procedure, which significantly improves the URLLC link adaptation accuracy. The proposed procedure applies a low-pass infinite impulse response (IIR) filtering of the measured interference, which effectively deals with the large load fluctuations due to the sporadic URLLC traffic.
- As there is no such thing as a free lunch, we determine the cost in terms of eMBB throughput for satisfying the stringent URLLC latency and reliability requirements.

An extensive system-level evaluation is carried out to quantify the benefits of the proposed enhancements. The presented results offer insight on the maximum offered URLLC traffic load that can be tolerated in the system, as well as its sensitivity to the URLLC-eMBB traffic composition, URLLC payload size, and link adaptation and scheduling setting. The complexity of our system model prevents a purely analytical evaluation without omitting many important practical aspects. The performance is therefore assessed via highly detailed system-level simulations, following the 5G NR evaluation methodology agreed in the 3rd Generation Partnership Project (3GPP) [2]. The simulator includes explicit and detailed modelling of the majority of radio resource management (RRM) functionalities, and link-to-system mapping for determining the error probability of each data transmission. These mechanisms, as well as the proposed techniques are based on underlying mathematical models, which are derived and described in the paper. When conducting the simulations, good practice in ensuring trustworthy and statistically-reliable results is applied.

The rest of the paper is organized as follows: Section II describes the considered network and traffic model, and the performance metrics. Section III outlines the RRM considerations, including the proposed radio resource scheduling and link adaptation enhancements. The simulation assumptions are outlined in Section IV. The performance results are shown in V, followed by concluding remarks in Section VI.

## II. SETTING THE SCENE

### A. NETWORK LAYOUT AND TRAFFIC MODEL

We follow the 5G NR modelling assumptions for a wide-area macro cellular scenario as outlined in [2]. This consists

of $C$ cells which are deployed in a sectorized manner, with three sectors per site and 500 meter inter-site distance. A set of $U$ UEs are uniformly distributed across the network area. Two different traffic compositions are considered: In case (i), the $U$ UEs are configured with URLLC type-of traffic. This consists of small payload sizes of $B$ Bytes (ranging from 32 to 200 Bytes [2]) that arrive for each URLLC UE in the downlink direction, following a Poisson arrival process with mean arrival rate $\lambda$ [payload/s]. This traffic model is known as FTP Model 3 in 3GPP [2]. Case (ii) consists of a mix of URLLC and eMBB UEs. eMBB UEs are modelled with background full buffer best-effort downlink traffic.

The following notation is used throughout the paper: set of cells and UEs are denoted by $\mathcal{C} = \{1, ..., C\}$ and $\mathcal{U} = \{1, ..., U\}$, respectively. To distinguish the UE type, we define $\mathcal{U}_{urllc} = \{1, ..., U_{urllc}\} \subseteq \mathcal{U}$ as the set of URLLC UEs, and $\mathcal{U} - \mathcal{U}_{urllc}$ as the set of eMBB UEs. We use superscript $\mathcal{U}^c$ to indicate the set of users connected to cell $c$. The *URLLC offered load*, defined as $L_{urllc} = U_{urllc} \cdot B \cdot \lambda / C$, is used to indicate the average amount of URLLC traffic that is offered per cell. The time domain is slotted into subframes or TTIs, each containing a set $\mathcal{P} = \{1, ..., P\}$ of physical resource blocks (PRB) in the frequency domain.

### B. FRAME STRUCTURE AND NUMEROLOGY

Users are dynamically multiplexed on a time-frequency grid of resources, using orthogonal frequency division multiple access (OFDMA) and frequency-division duplexing (FDD). The physical layer numerology follows the agreements in 3GPP for 5G NR: 15 kHz sub-carrier spacing (SCS), 14 OFDM symbols per 1 ms, and a PRB size of 12 sub-carriers (180 kHz) as the baseline configuration, although options with $2^N$ scaling of the SCS ($N \in [1, 2, ..., 5]$) are also allowed in the standard [2]. The carrier bandwidth configuration is 20 MHz, corresponding to $P = 100$ PRBs. The 3GPP has also agreed on using different TTI durations in accordance with the user-specific requirements. The possible time-domain scheduling resolutions include a *slot*, composed of 14 OFDM symbols; and *mini-slots* of 1 to 13 OFDM symbols [2]. For the purpose of achieving low latency, we assume that URLLC and eMBB UEs are scheduled on a 2 OFDM symbol (0.143 ms) mini-slot resolution. We refer to [15], [17], and [27] for URLLC and eMBB system-level performance results with different TTI durations.

Each data transmission to a user is indicated with a scheduling grant. The scheduling grant contains information on the specific time-frequency resource allocation for each user, the employed MCS, and other transmission parameters required to decode the data. In line with [28], the control channel (CCH) for transmitting the scheduling grant is accommodated within the resources assigned to each user (i.e. in-resource CCH). The coding rate of the in-resource CCH is dynamically adapted in accordance with the user's channel condition, as expressed in the CQI report. We assume that the in-resource CCH will carry similar information as the LTE physical downlink control channel (PDCCH), and we therefore use the PDCCH link-level performance [29] as a reference. That is, a minimum of 36 resource elements (REs) in order to transmit the CCH with a BLEP of 1%, with additional repetition encoding in form of aggregation levels 2, 4, 8, depending on the user's channel condition. One RE corresponds to one OFDM subcarrier symbol. Note that the considered in-resource CCH allows for more flexible scaling of the control channel overhead, as compared to LTE where the PDCCH overhead is either 7%, 14%, or 21% [30].

On each scheduling opportunity, the resource allocation to a user must be sufficiently large to accommodate the in-resource CCH as well as a reasonable data payload and reference symbols [27]. Table 1 summarizes the required number of REs for the CCH depending on the user-specific signal to interference and noise ratio (SINR), as well as the corresponding minimum frequency-domain allocation size assumed in this work.

**TABLE 1. CCH overhead and scheduling format for a 2-symbol (0.143 ms) TTI size [29].**

| SINR [dB] | In-resource CCH overhead | Frequency-domain minimum allocation size |
|---|---|---|
| $(-\infty, -2.2)$ | $8 \cdot 36 = 288$ REs | 14 PRBs (336 REs) |
| $[-2.2, 0.2)$ | $4 \cdot 36 = 144$ REs | 8 PRBs (192 REs) |
| $[0.2, 4.2)$ | $2 \cdot 36 = 72$ REs | 5 PRBs (120 REs) |
| $[4.2, \infty)$ | $1 \cdot 36 = 36$ REs | 3 PRBs (72 REs) |

### C. LATENCY BUDGET

For each UE, data from higher layers are received at the serving cell and stored in a user-specific transmission buffer as illustrated in Fig. 1. The URLLC latency is measured from the moment a URLLC payload arrives at the serving cell until it is successfully received at the UE. Assuming a first transmission error probability of $P_e$, the latency $T$ for a successfully received first transmission equals $\tau$, expressed as $P(T = \tau) = 1 - P_e$, with

$$\tau = \max(\tau_q; \tau_f) + \tau_{tx} + \tau_{prx}, \quad (1)$$

where $\tau_q$ is the queuing delay of the URLLC payload at the base station, $\tau_f$ is the so-called frame alignment, $\tau_{tx}$ is the transmission time of the payload, and $\tau_{prx}$ is the processing time at the receiver end. The receiver processing time is constant. The frame alignment is uniformly distributed between zero and the TTI length $\tau_{tti}$, i.e. $\tau_f \in U(0, \tau_{tti})$. The transmission of a URLLC payload takes at least one TTI but may also take multiple TTIs depending on the available resources, payload size, and radio channel conditions, i.e. $\tau_{tx}$ is a discrete random variable with probability mass function fulfilling $\sum_{N=1}^{\infty} N \cdot \tau_{tx} = 1$. The base station queuing delay $\tau_q$ accounts for the time the data arrives at the serving cell until it is considered for scheduling (transmission). The value of $\tau_q$ depends on the arrival rate of payloads at the base station, the scheduling policy, as well as on how

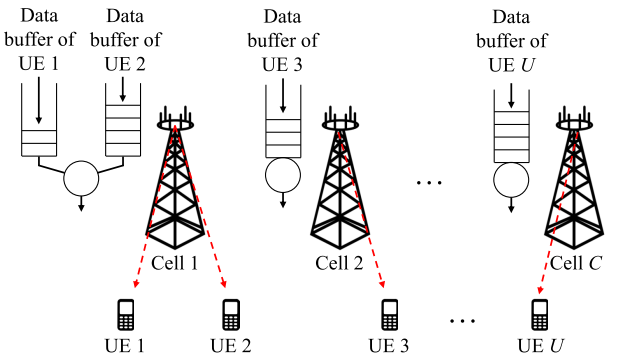


**FIGURE 1.** System model.

fast the payloads are transmitted, and thus it is also a random variable. For networks with low offered load, $\tau_q \to 0$, as the incoming payloads are transmitted much faster than they arrive. As the offered load increases, queuing theory allows to determine the distribution of $\tau_q$ under typically simplified assumptions, e.g. fixed payload transmission time at the base station (constant air interface capacity) or fixed amount of concurrent payload transmissions per TTI [16]. However, for a realistic radio system, the distribution of both $\tau_q$ and $\tau_{tx}$ are very hard to analytically derive, as the rate at which payloads are successfully transmitted (and received by the UEs) is a multi-dimensional random process, depending on the users' experienced SINR condition, and therefore also on the time-variant other-cell interference and radio propagation conditions.

For cases where the first transmission is erroneously decoded by the UE, a HARQ retransmission is triggered. The decoding error probability of a first HARQ retransmission is denoted $P'_e$, where it can be assumed that $P'_e \ll P_e$ due to the HARQ soft combing gain. We can therefore express

$$P(T = \tau + \tau_{HARQ}) = P_e \cdot (1 - P'_e), \tag{2}$$

where $\tau_{HARQ}$ denotes the HARQ round trip time (RTT). Note that (2) assumes that HARQ retransmissions are always prioritized, and hence are not subject to queuing delays. In line with [15], we assume $\tau_{HARQ} = 4 \cdot \tau_{tti}$. Fig. 2 shows an example time-line of the transmission of a URLLC payload. For a matter of simplicity, we assume $\tau_{tx} = \tau_{tti} = 0.143$ ms and $\tau_q = 0$. Under these conditions, the maximum latency corresponds to $6 \cdot 0.143$ ms = 0.86 ms, hence satisfying the 1 ms latency target. In practice, $\tau_q \geq 0$ for each URLLC transmission, and its effect will be closely analysed when presenting the performance results in Section V.

## III. RADIO RESOURCE MANAGEMENT CONSIDERATIONS

### A. BASIC LINK ADAPTATION

The packet scheduler and link adaptation functionality play an important role in fulfilling the users' QoS requirements. Dynamic link adaptation is assumed for all users (regardless of the service type) by adjusting the used MCS per downlink

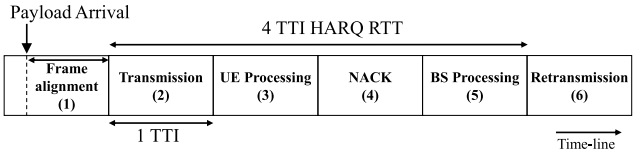


**FIGURE 2.** Diagram of URLLC downlink data transmissions, assuming one HARQ retransmission and $\tau_q = 0$.

transmission. The supported MCS index, $m$, for a certain user is expressed as,

$$m = \arg\max_m \{R_m | P_e \leq P_{target}\}, \tag{3}$$

corresponding to the largest supported data rate, $R_m$, by using MCS index $m$ without exceeding a block error probability of $P_{target}$. The default setting for the BLEP target is $P_{target} = 0.1$ (as also assumed for LTE CQI reporting [30]). In practice, this is achieved by having the UEs measure the experienced SINR, followed by evaluation of (3), given knowledge of the BLEP vs SINR mapping curve for each of the supported MCSs. However, due to UE SINR estimation imperfections and CQI reporting delays (during which the SINR conditions may change), the well-known outer loop link adaptation (OLLA) algorithm is applied, where the received CQI values are offset by a certain factor $O$ (a.k.a. the OLLA offset) [31]. Thus, in effect, the offset $O$ is subtracted from the estimated SINR before selecting the MCS. This is feasible since the CQI table is designed to have constant SINR offset between the entries. Thus, if the SINR offset between the CQI table entries is 1 dB, and the OLLA offset equals $X$ dB, then the received CQI index is offset by round($X$) steps before being used. In line with [31], the factor $O$ is increased by $\Delta_{up}$ upon receiving a NACK from a previous transmission, while it is decreased by $\Delta_{down}$ if receiving a ACK. Thereby, the long-term average block error probability is controlled to converge to $\bar{P}_e = (1 + \Delta_{up}/\Delta_{down})^{-1}$. The MCS for the eMBB users is adjusted to reach $\bar{P}_e = 0.1$ (10%). For the URLLC UEs, the BLEP target should be sufficiently low to fulfil their reliability and latency requirements, but also not lower than that as this would mean using unnecessary transmission resources (i.e. PRBs), potentially harming other users in the system [14], [15]. In the following, we first describe the conventional resource allocation approach, where the link adaptation for URLLC users is configured to achieve a static average long term block error probability of $\bar{P}_e = 0.01$ (1%) or $\bar{P}_e = 0.001$ (0.1%). Next, we present an improved resource allocation method that dynamically adjusts the BLEP of each individual URLLC transmission in accordance with the instantaneous experienced load per cell.

### B. RESOURCE SCHEDULING WITH FIXED BLEP TARGET

The baseline resource allocation scheme works as follows: on each TTI $n$, each cell $c \in \mathcal{C}$ independently allocates up to $P$ PRBs to its associated users $\mathcal{U}^c$, taking into account the

user-specific QoS class and radio-channel conditions [31]. The QoS-awareness is achieved by dividing the scheduling procedure into two stages: a first one where PRBs are allocated to the URLLC users with pending data transmission[1]; and a second one where the remaining PRBs (if any) are allocated to the $\mathcal{U}^c - \mathcal{U}^c_{urllc}$ eMBB UEs (i.e. hard priority type-of scheduling). For each step, each PRB $p$ is assigned to the user $u^*$ that maximizes the well-known *proportional fair* (PF) metric, i.e.,

$$u_p^* = \arg\max_u \left\{ \frac{r_{u,p}[n]}{T_u[n]} \right\}, \tag{4}$$

where $n$ is the discrete time index for the scheduling interval, $r_{u,p}$ is an estimate of the instantaneous supported data rate of user $u \in \mathcal{U}^c$ in the $p$-th PRB, and $T_u$ is its average delivered user throughput in the past. The value of $r_{u,p}$ is estimated based on the periodical frequency-selective CQI report sent by each UE, whereas $T_u[n]$ is calculated recursively using a moving average filter and is only updated for users that have data buffered [32]. The use of the scheduling metric in (4) is especially relevant when multiple URLLC UEs need to be scheduled in the same TTI, as it implicitly captures the frequency-varying channel quality. The cell index has been left out of (4) for the sake of simplicity.

On each scheduling interval, the allocation size (i.e. number of PRBs) to a user can be as small as indicated in Table 1, whereas the maximum allocation size is a function of the available resources, user pending data, and the employed MCS. The MCS is selected in order to reach a fixed BLEP target ($\bar{P}_e = 0.1$ for eMBB, and $\bar{P}_e = 0.01$ or $\bar{P}_e = 0.001$ for URLLC).

### C. RESOURCE SCHEDULING WITH DYNAMIC BLEP ADJUSTMENT

Fig. 3 outlines the operation of the proposed resource allocation scheme with dynamic BLEP adjustment for the URLLC transmissions. The scheduling procedure consists of three steps. Steps 1 and 3 are inherited from Section III-B. In step 1, each cell $c$ allocates PRBs to its associated URLLC UEs based on their experienced channel quality as expressed in the CQI report. The key point in this step is that each URLLC UE $u$ with pending data transmission receives an allocation of size $x_u$ PRBs ($x_u \in [0, 1, ..., P]$ and $\sum_{u \in \mathcal{U}^c} x_u \leq P$), such that the URLLC payload can be transmitted with a modest initial BLEP target, e.g. $\bar{P}_e = 0.01$. Once the initial $X^c = \sum_{u \in \mathcal{U}^c} x_u$ PRBs have been allocated, step 2 consists on assigning a proportion $\Gamma$ ($0 < \Gamma < 1$) of the remaining $P - X^c$ PRBs to the already allocated URLLC users. The additional resources will allow to transmit the URLLC small payload with a more conservative MCS (i.e. even further reduced BLEP). The question is now: how to select and distribute the $\Gamma \cdot (P - X^c)$ PRBs among the different URLLC users? We assume that each URLLC user $u$ is allocated with $\gamma_u$ additional PRBs,

[1]HARQ retransmissions are prioritized over new transmissions in line with [31].

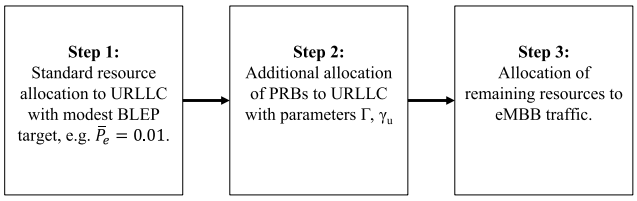


**FIGURE 3.** Scheduling procedure with dynamic BLEP adjustment.

where $\gamma_u$ is calculated proportionally to the user's initial allocation size $x_u$, i.e.,

$$\gamma_u = \frac{\Gamma \cdot (P - X^c)}{X^c} \cdot x_u \ . \tag{5}$$

As an example, for $\Gamma = 0.5$, two URLLC users who receive an initial allocation size $x_1 = 16$ PRBs, and $x_2 = 21$ PRBs, would be scheduled on $\gamma_1 = 0.5 \cdot (100 - 37)/37 \cdot 14 \approx 12$ and $\gamma_2 = 0.5 \cdot (100 - 37)/37 \cdot 21 \approx 18$ additional PRBs in step 2. The $\gamma_u$ additional PRBs for each user are selected following the PF rule as described in (4).

In step 3, the remaining $(1 - \Gamma) \cdot (P - X^c)$ PRBs are allocated to the eMBB UEs. Note that for $\Gamma = 1$, eMBB users will only be scheduled on TTIs where no URLLC traffic is present. This setting provides the highest URLLC reliability, at the expense of the largest eMBB throughput degradation; whereas the case with $\Gamma = 0$ corresponds to the baseline scheduling operation (as described in Section III-B). In essence, the proposed resource allocation technique aims at scheduling URLLC UEs with a modest BLEP target, e.g. $\bar{P}_e = 0.01$ for URLLC, but can be lower depending on the configured $\Gamma$, and the experienced URLLC traffic load on each scheduling instant.

### D. ACCURATE CQI MEASUREMENTS

As mentioned, the UE selects a CQI based on the experienced channel quality. The post-receiver SINR is typically used for such purpose as it captures the potential interference cancellation/suppression capabilities of the UE receiver. For single-stream transmissions, as considered in this work, a common expression for the post-receiver SINR experienced by user $u$ in a system with $N_r$ receive antennas at the UE, and $N_t$ transmit antennas at the cells is [33],

$$\Psi_{u,c} = \frac{\Omega_{u,c} \|\mathbf{g}_u \mathbf{H}_{u,c} \mathbf{f}_c\|^2 \ P_c}{\sum_{i \in \mathcal{I}} \Omega_{u,i} \|\mathbf{g}_u \mathbf{H}_{u,i} \mathbf{f}_i\|^2 P_i + \sigma^2_{n,u}}, \tag{6}$$

where sub-index $c$ denotes the serving cell; $\mathcal{I} \subseteq \mathcal{C}$ is the set of cells that create interference to user $u$; $\Omega_{u,i}$ denotes the large scale fading (pathloss and shadowing); $\mathbf{g}_u \in \mathbb{C}^{1 \times N_r}$ is the receiver filter; $\mathbf{H}_{u,i} \in \mathbb{C}^{N_r \times N_t}$ represents the small scale fading; $\mathbf{f}_i \in \mathbb{C}^{N_t \times 1}$ is the transmit precoder; $P_i$ is the transmit power; and $\sigma^2_{n,u}$ is the total background noise power received by the user.

The nominator in (6) represents the power of the desired-signal and it is measured using the cell-specific reference signals transmitted by the serving cell [30]. In contrast,

**TABLE 2. Simulation assumptions.**

| Parameter | Value |
|---|---|
| Network environment | 3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance [2] |
| Carrier configuration | 20 MHz carrier bandwidth at 2 GHz |
| Propagation | $128.1 + 37.6\log(R[km])$ dB; Log-Normal shadowing with 8 dB standard deviation |
| PHY numerology | 15 kHz subcarrier spacing; 12 subcarriers per PRB; TTI size of 2 OFDM symbols (0.143 ms) |
| Control channel | Error-free in-resource scheduling grants with dynamic link adaptation [28] |
| Data channel MCS | QPSK to 64QAM, with same coding rates as in LTE |
| CSI | LTE-alike CQI and PMI, reported every 5 ms; Sub-band size: 8 PRBs |
| | CQI filtering ($\alpha = 0.01$) for cases with only URLLC traffic |
| Reference signals overhead | 4 resource elements per PRB |
| Antenna configuration | 2 x 2 single-user single-stream MIMO with LTE-alike precoding and MMSE-IRC receiver |
| Packet scheduler | Proportional Fair with priority for URLLC traffic and different $\Gamma$ settings |
| Block error probability target | eMBB: $\bar{P}_e = 0.1$; URLLC: $\bar{P}_e = 0.01$ or $\bar{P}_e = 0.001$ with different $\Gamma$ settings |
| HARQ | Asynchronous HARQ with Chase combining and 4 TTI RTT; Max. 6 HARQ retransmissions |
| RLC | RLC Unacknowledged mode |
| Traffic composition | Case a) 210 URLLC UEs; Case b) 210 URLLC UEs + 105 eMBB UEs |
| UE distribution | Uniformly distributed in outdoor locations; No mobility |
| Traffic model | URLLC: FTP3 downlink traffic with 32, 50 or 200 Byte payload size; eMBB: full buffer |
| URLLC Offered load | 1 - 8 Mbps average load per cell |

the received interference and noise (denominator) is typically measured on radio resources used for data transmission, meaning that it captures the time and frequency transmit-power variations in accordance to the instantaneous load at each interfering cell. Accurate link adaptation is therefore challenging in scenarios with only URLLC traffic: Due to the relatively small payloads, a URLLC transmission generally occupies a subset of the available PRBs within a TTI. This fact, together with the sporadic nature of URLLC traffic, result in a rapidly changing interference pattern, hence making it difficult to accurately select and report a CQI that fulfils the specified BLEP constraint upon the downlink transmission. This problem is also well-known from LTE system-level performance analyses in non-fully loaded networks [34]. However, it is exacerbated in this scenario as drastic variations occur from TTI to TTI and on a PRB basis.

In LTE, the UE determines the CQI based on a finite number of channel quality measurements obtained from a relatively short measuring window [30]. Our proposal is to modify the UE measurement procedure of the CQI report, by including historical information of the experienced interference. On each TTI $n$, each UE $u$ measures the interference with a certain PRB resolution (a.k.a. sub-band). The instantaneous interference measurement on the $s$-th sub-band is given by,

$$y_{u,s}[n] = \sum_{i \in \mathcal{I}} \Omega_{u,i}[n] \|\mathbf{g}_u[n]\mathbf{H}_{u,i,s}[n]\mathbf{f}_i[n]\|^2 P_{i,s}[n] + \sigma^2_{n,u}. \tag{7}$$

Note that (7) is basically the denominator of (6), but limited to the $n$-th TTI and $s$-th sub-band. Each $y_{u,s}[n]$ measurement is filtered with a low-pass first-order IIR filter, resulting in the following smoothed value:

$$s_{u,s}[n] = \alpha \cdot y_{u,s}[n] + (1-\alpha) \cdot y_{u,s}[n-1], \tag{8}$$

where $\alpha$ is the forgetting factor (FF) of the filter ($0 < \alpha < 1$). The per-sub-band CQI, which is periodically reported to the serving cell, contains the low-pass filtered interference information $s_{u,s}[n]$ together with the latest desired-signal fading information, i.e.:

$$\Psi_{u,c,s}[n] = \frac{\Omega_{u,c}[n]\|\mathbf{g}_u[n]\mathbf{H}_{u,c,s}[n]\mathbf{f}_c[n]\|^2 P_c}{s_{u,s}[n]}, \tag{9}$$

Note that the received power from the serving cell (nominator of (9)) 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 [30]. The FF $\alpha$ determines how much weight is given to the latest interference measurement as compared to the previous ones. Following the previous work in [15], we use $\alpha = 0.01$, which provides significantly latency and reliability improvement.

## IV. SIMULATION ASSUMPTIONS

The performance evaluation is based on dynamic system-level simulations following the 3GPP 5G NR methodology [2]. The default simulation assumptions are summarized in Table 2. The network layout, UE distribution and traffic follow the description presented in Section II-A. The network is composed of $C = 21$ cells, where $U_{urllc} = 210$ URLLC UEs are uniformly distributed (10 UEs per cell in average). eMBB background traffic is modelled with 105 additional UEs (5 UEs per cell in average) with best-effort full-buffer downlink traffic.

The simulator's time-resolution is one OFDM symbol, and it includes explicit modelling of the radio resource management functionalities described in Section II and III.

**TABLE 3.** URLLC RU[%] for different URLLC offered loads and URLLC payload sizes.

| $L_{urllc}$ [Mbps] | 200 Byte Payload | | 50 Byte Payload | | 32 Byte Payload | |
|---|---|---|---|---|---|---|
| | w/o eMBB | w/ eMBB | w/o eMBB | w/ eMBB | w/o eMBB | w/ eMBB |
| 1 | 1.6 | 3.9 | 2.2 | 4.6 | 2.8 | 5.3 |
| 2 | 3.3 | 7.9 | 4.5 | 9.3 | 5.6 | 10.7 |
| 4 | 6.9 | 16.3 | 9.3 | 18.9 | 11.7 | 21.4 |
| 6 | 11.3 | 24.4 | 14.4 | 28.6 | 17.9 | 31.8 |
| 8 | 16.6 | 33.2 | 20.0 | 38.2 | 24.7 | 43.2 |

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 member companies. The basic methodology is outlined in the following: 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 [35]. The MMSE-IRC receiver is modelled with the following receiver filter expression,

$$\mathbf{g}_u = \mathbf{f}_c^H \mathbf{H}_{u,c}^H \mathbf{R}^{-1}, \tag{10}$$

where $\mathbf{R}$ is the interference covariance matrix, i.e.,

$$\mathbf{R} = P_c \mathbf{H}_{u,c} \mathbf{f}_c \mathbf{f}_c^H \mathbf{H}_{u,c}^H + \sum_{i \in \mathcal{I}} P_i \mathbf{H}_{u,i} \mathbf{f}_i \mathbf{f}_i^H \mathbf{H}_{u,i}^H + \sigma_{n,u}^2 \mathbf{I}, \tag{11}$$

where $\mathbf{I}$ is the identity matrix. Given the SINR per RE, the effective exponential SINR model [36] is applied for link-to-system-level mapping to determine if the transmission was successfully decoded. Asynchronous adaptive HARQ with Chase Combining is applied in case of failed transmissions, and the SINRs for the different HARQ transmissions are linearly added [37]. The maximum number of HARQ retransmissions is limited to 6 for both URLLC and eMBB UEs; although, in practice, URLLC transmissions experience at most two HARQ retransmissions due to the low initial BLEP target. Closed-loop single-stream single-user 2x2 MIMO transmission mode is assumed, i.e. benefiting from both transmission and reception diversity against fast fading radio channel fluctuations [8]. The use of closed-loop MIMO schemes provide a valuable SINR gain over open-loop schemes, even under the presence of errors in the feedback channel [38]. Dynamic link adaptation is applied for both data and the in-resource CCH, based on periodical frequency-selective CQI reports from the UEs. The simulator does not consider user mobility; however, the dynamic traffic model and fast fading effects (calculated for a UE speed of 3 km/h) provide significant variability to the channel conditions. Unless otherwise mentioned, we assume low-pass IIR CQI measurements at the UE (Section III-D) for cases with only URLLC traffic. Each cell independently schedules its users with full priority for URLLC traffic.

For each URLLC UE, payloads of $B$ Bytes are generated in the downlink direction following a Poisson distribution with arrival rate $\lambda$. Payload sizes of 32 Bytes, 50 Bytes and 200 Bytes are considered [2]. When presenting the results, we will refer to *URLLC offered load* or simply *offered load* to indicate the average amount of URLLC traffic that is offered per cell.

The latency (defined in Section II-C) of each successfully received URLLC payload is collected and used to form empirical complementary cumulative distribution functions (CCDF). For URLLC, the key performance indicator (KPI) is the achievable latency with 99.999% probability, i.e. the $10^{-5}$ percentile of the CCDF. For eMBB users, the primary KPI is the 5th percentile and 50th percentile (or median) of the downlink end-user throughput.

The simulation time corresponds to at least 5 million successfully received URLLC payloads. Assuming that the obtained latency samples are uncorrelated, this allows to estimate the 99.999% percentile of the latency cumulative distribution with an error margin of at most $\pm 5\%$,[2] with a 95% confidence level [39]. In practice, some correlation is present among the latency samples, slightly increasing the error margin of the results.

## V. PERFORMANCE RESULTS

The URLLC latency and eMBB throughput performance is evaluated under different offered load conditions, scheduling policies, and URLLC payload sizes. The percentage of PRBs allocated to URLLC traffic (we refer to this as URLLC resource utilization (RU)) is summarized in Table 3. Here, we assume a fixed BLEP target $\bar{P}_e = 0.001$ (0.1%) for URLLC traffic and $\Gamma = 0$. As expected, the URLLC RU increases with the URLLC offered load $L_{urllc}$. This growth is non-linear for cases without eMBB traffic. Apart from the larger volume of data that needs to be delivered, higher $L_{urllc}$ results in larger inter-cell interference and consequently lower signal quality for URLLC users. In contrast, cases with full-buffer eMBB traffic correspond to a fully loaded network. This results in close-to linear increase of the URLLC RU vs $L_{urllc}$, since the signal quality of the UEs does not change with the offered load.

### A. URLLC PERFORMANCE

We first focus on cases with 200 Byte URLLC payloads. Fig. 4(a) shows the CCDF of the URLLC latency under different offered load conditions, for the scenario without eMBB traffic. At the $10^{-5}$ percentile, it is observed that

[2]The actual error margin depends on the steepness of the latency cumulative distribution at the percentile of interest.

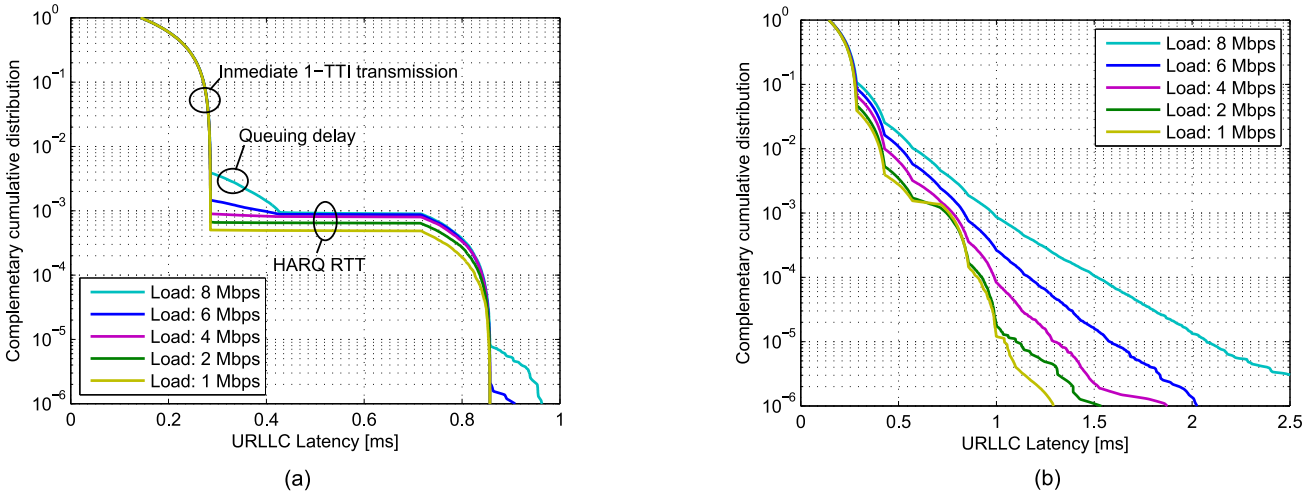


**FIGURE 4.** URLLC latency distribution for different URLLC offered load conditions, and traffic configurations. $\bar{P}_e = 0.001$ for URLLC; $\Gamma = 0$; 200 Byte payload.

the 1 ms latency requirement is fulfilled for all the considered offered loads of URLLC traffic. Specifically, a latency of ∼0.86 ms is achieved, matching the latency budget in Fig. 2. The different components contributing to the URLLC latency are also depicted. The upper part of the distribution ($10^0 - 10^{-3}$ percentile) represents the case where the URLLC payloads are immediately scheduled and correctly received at the UE. With $10^{-3}$ probability (equivalent to the URLLC-specific 0.1% BLEP target), the initial URLLC transmissions are not correctly received at the UE side. This triggers a HARQ retransmission, which is immediately scheduled after receiving the HARQ NACK at the serving cell. In addition to this, some temporary queuing delay is experienced at the cells' buffers when operating at a offered load of 8 Mbps, hence degrading the URLLC latency.

Fig. 4(b) shows the URLLC latency distribution for cases with eMBB traffic. It is observed that the 1 ms latency with $1 - 10^{-5}$ reliability is not fulfilled, even at low URLLC offered loads. Even though URLLC transmissions are fully prioritized by the packet scheduler, the larger inter-cell interference from scheduling eMBB users significantly degrades the URLLC latency performance. This is highlighted in Fig. 5 which shows the distribution of the instantaneous post-detection SINR of the URLLC users. Cases without eMBB traffic experience a 7 dB SINR degradation at the median when increasing the offered load from 1 Mbps to 8 Mbps. For the scenario where the network is fully loaded with eMBB traffic, the SINR is independent of the URLLC offered load, and significantly worse than the cases with only URLLC traffic. Lower SINR results in lower MCS for data transmissions and higher CCH overhead. As a consequence, larger amount of PRBs (see Table 3) is required to deliver the URLLC payloads, having a negative impact on the queuing delay at the cell and transmission delay (e.g. a URLLC payload not fitting in a single TTI).

The URLLC latency and reliability performance is tightly related to the scheduling and link adaptation settings.

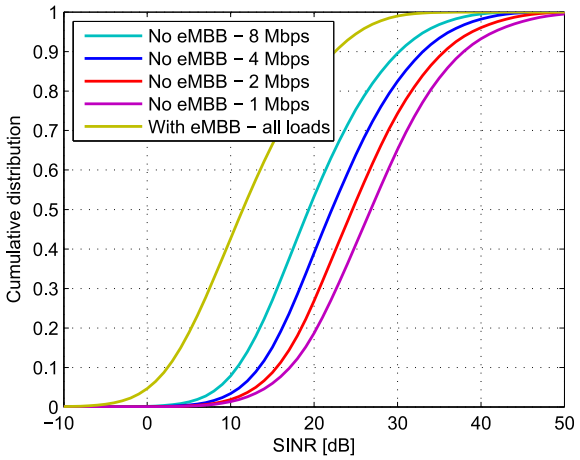


**FIGURE 5.** Instantaneous per-user per-subcarrier SINR for different URLLC offered load conditions, and traffic configurations. $\bar{P}_e = 0.001$ for URLLC; $\Gamma = 0$; 200 Byte payload.

Fig. 6(a) summarizes the latency performance at the $1 - 10^{-5}$ percentile for different offered traffic loads. For the scenario without eMBB traffic, we show settings with $\alpha = 0.01$ and $\alpha = 1$ of the low-pass CQI filtering enhancement.[3] For the case where eMBB traffic is present, we include different configurations of the scheduling technique ($\Gamma > 0$) presented in Section III-C. It is observed that the proposed CQI filtering scheme provides large gains, as low-pass information of the experienced interference is implicitly included in the CQI report. These benefits are especially relevant at high load when the cell activity is higher and more sporadic interference is experienced across the network. For cases with eMBB traffic and $\Gamma = 0$, the configured URLLC BLEP target has a large impact on the achievable latency. At low URLLC offered load, it is advantageous

[3]Note that this technique is less relevant in scenarios where eMBB background traffic is present, as stable full-load interference conditions are experienced.

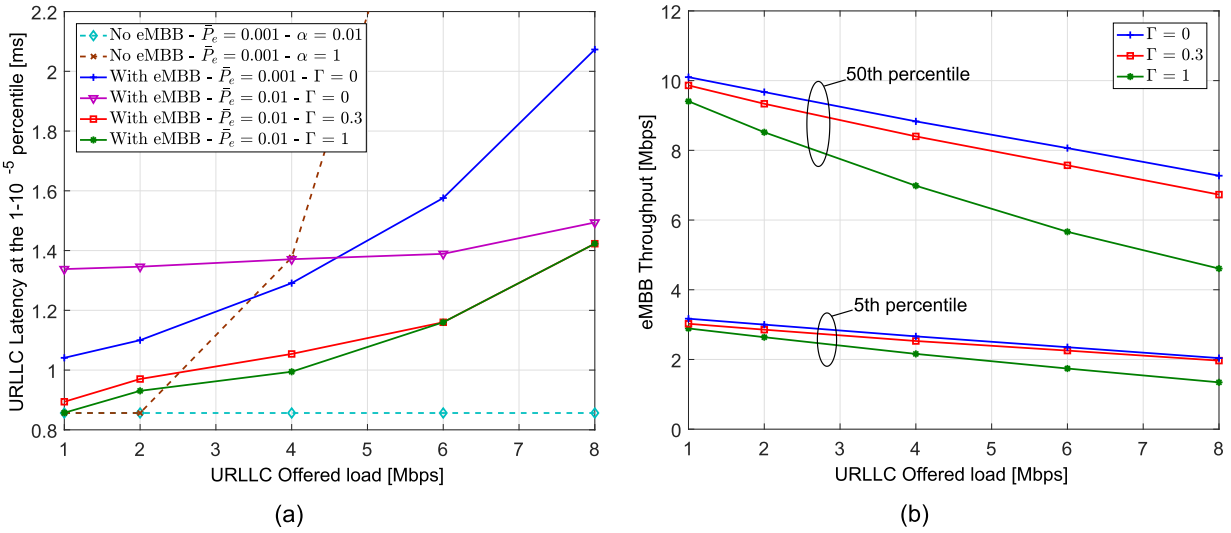


**FIGURE 6.** Summary of (a) URLLC latency at the $1 - 10^{-5}$-percentile, and (b) eMBB throughput at the 5th- and 50th-percentile. 200 Byte payload.

to operate with $\bar{P}_e = 0.001$ (0.1%) in order to reduce the occurrence of HARQ retransmissions (and the corresponding HARQ processing delay). As the load increases, non-negligible queuing starts to occur at the cells' buffers, which deteriorates considerably the latency performance. Under such circumstances, it is beneficial to operate with $\bar{P}_e = 0.01$ (1%) in order to increase the spectral efficiency of the system and reduce the queue length. Due to these tradeoffs, the proposed resource allocation algorithm ($\Gamma > 0$) provides much better latency performance. Recall from Section III-C that the proposed scheduling technique aims at scheduling URLLC UEs with a BLEP target of at most 1%, but can be lower depending on the instantaneous URLLC load at each cell. The 1 ms URLLC latency requirement is fulfilled for offered loads up to 2 Mbps, although the latency performance at higher URLLC offered loads is still decent ($\leq$1.4 ms). Settings with $\Gamma = 1$ provide the best URLLC performance, whereas $\Gamma = 0.3$ still provide relevant latency improvement with only minor impact on the eMBB throughput performance, as will be presented next.

### B. EMBB PERFORMANCE

Fig. 6(b) shows the 5th- and 50th- percentile of the eMBB throughput under different scheduling and traffic settings. As expected, the eMBB throughput decreases as we increase the URLLC offered load. Configurations with $\Gamma = 0$ achieve the highest eMBB throughput, whereas cases with $\Gamma > 0$ experience lower throughput at the expense of reduced URLLC latency, as shown in Fig. 6(a). Particularly, the setting with $\Gamma = 0.3$ offers significant latency reduction (e.g. from $\sim$1.3 ms down to $\sim$1.05 ms at 4 Mbps offered load) with a small throughput degradation ($\leq$10% for any URLLC offered load condition). Furthermore, the proposed solution is highly robust and flexible, as it brings relevant URLLC latency and reliability improvements for a wide range of offered load conditions without requiring fine adjustment of the link adaptation settings.

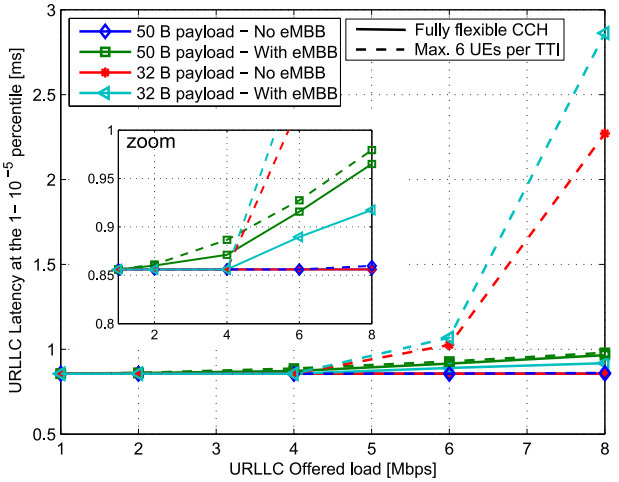


**FIGURE 7.** URLLC latency performance for different payload sizes and CCH settings. $\bar{P}_e = 0.001$ for URLLC; $\Gamma = 0$.

### C. SENSITIVITY TO THE URLLC PAYLOAD SIZE

The sensitivity to the URLLC payload size is presented next. As observed in Table 3, settings with smaller payload sizes experience larger RU as compared to the 200 Byte payload case, which is a consequence of the larger CCH overhead. Under these circumstances, the considered in-resource CCH brings relevant benefits, as it allows more flexible scaling of the CCH as compared to LTE. Fig. 7 shows the URLLC latency at the $1 - 10^{-5}$ percentile for 32 Byte and 50 Byte payload sizes, with and without eMBB traffic. In order to illustrate the benefits of the in-resource CCH, we include cases where the maximum number of scheduled URLLC UEs per TTI is limited to six.[4] It is observed that the enforced CCH restriction considerably degrades the URLLC performance.

[4]Common assumption for LTE performance evaluation, given the limited PDCCH capacity [30].

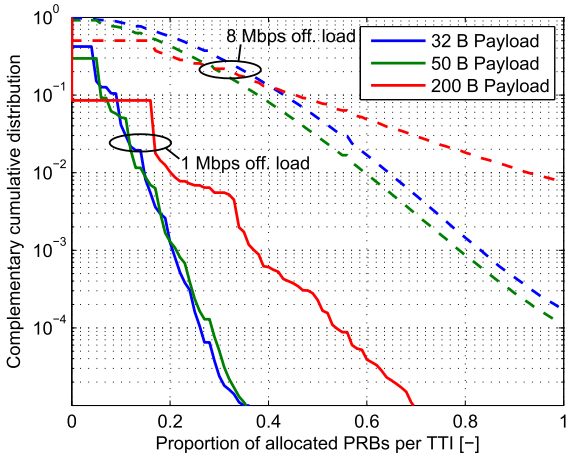


**FIGURE 8.** Proportion of allocated PRBs per TTI per cell for different URLLC payload sizes and offered loads. No eMBB traffic. $\bar{P}_e = 0.001$ for URLLC; $\Gamma = 0$.

As an example, the setting with 32 Byte payload does not achieve the 1 ms latency requirement for $L_{urllc} \geq 6$ Mbps, even for cases without eMBB traffic. This degradation is mainly due to the enforced CCH restriction, meaning that radio resources are left unused due to the limited CCH capacity. Some performance degradation is also observed for settings with 50 Byte payload size; although the 1 ms latency requirement is still fulfilled. In contrast, cases without the enforced CCH restriction (fully flexible CCH) achieve significantly better performance. The URLLC requirements are fulfilled from low load to high load, also in cases where eMBB traffic is present in the network. Such good performance is mainly due to the lower amount of PRBs required for the transmission of smaller payloads. This is reflected in Fig. 8, where the empirical distribution of the allocated PRBs per TTI for 1 and 8 Mbps offered load is shown for the scenario with only URLLC traffic. Even though small payload sizes experience larger average RU (Table 3), the variance of the allocated PRBs per TTI is much larger for cases with a payload size of 200 Bytes. As an example, for cases with 8 Mbps and 200 Byte payload, the 100% of the PRBs are scheduled with a non-negligible $10^{-2}$ probability. This results in large queuing and transmission delay, meaning that URLLC transmissions are not immediately scheduled upon arrival. This temporary queuing occurs less often for payload sizes of 32 Bytes or 50 Bytes, which explains the much better latency performance (for cases without CCH limitations).

Although not shown in the results, similar findings can be obtained by fixing the URLLC payload size and varying the available carrier bandwidth. For instance, doubling the bandwidth would allow to increase by a factor of two or more the URLLC load that can be tolerated in the system [16]. Hence, increasing the available bandwidth, or operating with small payload sizes are two relevant approaches to improve the resource efficiency of the system, i.e. allowing high resource utilization while still achieving the stringent URLLC requirements.

## VI. CONCLUSIONS

In this paper, we have presented solutions for efficient multiplexing of URLLC and eMBB traffic on a shared channel. Specifically, a dynamic resource allocation technique have been proposed which provides a simple, yet effective method to determine how the radio resources should be distributed between the two service classes, in accordance with the well-known tradeoffs between reliability, latency and spectral efficiency. A detailed system-level analysis of the URLLC and eMBB downlink performance shows significant gains of the proposed solution, reducing the URLLC latency from 1.3 ms to 1 ms at the 99.999% percentile, with less than 10% degradation of the eMBB median throughput performance, as compared to conventional scheduling techniques. The main messages brought by this paper are the following: (i) It is possible to multiplex URLLC with eMBB traffic such that the URLLC requirements are fulfilled even under very-high interference from serving the eMBB users. (ii) There is a price to pay in terms of eMBB throughput performance in order to achieve stringent latency and reliability requirements of URLLC. And (iii) the URLLC performance is highly sensitive to the traffic characteristics, particularly the relation between the available carrier bandwidth, the offered load, and the URLLC payload size. As an example, cases with relatively large URLLC payload size (200 Bytes) fulfil the requirements only for low or medium offered load conditions ($<$ 4 Mbps), whereas settings with smaller payload size (32-50 Bytes) can operate at higher load ($\geq$8 Mbps). In the latter case, we have highlighted the importance of a flexible control channel design in order to avoid problems of control channel blocking, as known from LTE.

Future work must consider a more realistic modelling of eMBB traffic, e.g. finite buffer traffic including the transmission control protocol (TCP) flow control mechanisms. Also, accounting for control channel errors and ACK/NACK misdetections is of relevance to further assess the URLLC latency and reliability performance.

## REFERENCES

[1] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document 3GPP TR 38.913 v14.1.0, Jan. 2017.
[2] *Study on New Radio Access Technology Physical Layer Aspects*, document 3GPP TR 38.802 v14.0.0, Mar. 2017.
[3] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. Int. Conf. 5G Ubiquitous Connectivity*, Nov. 2014, pp. 146–151.
[4] A. Frotzscher *et al.*, "Requirements and current solutions of wireless communication in industrial automation," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Jun. 2014, pp. 67–72.
[5] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen, and P. Mogensen, "Automation for on-road vehicles: Use cases and requirements for radio design," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2015, pp. 1–5.
[6] V. C. Gungor *et al.*, "A survey on smart grid potential applications and communication requirements," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 28–42, Feb. 2013.
[7] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. P. Fettweis, "5G-enabled tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.
[8] 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*, Dec. 2015, pp. 1–6.

[9] F. Kirsten, D. Öhmann, M. Simsek, and G. P. Fettweis, "On the utility of macro- and microdiversity for achieving high availability in wireless networks," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug./Sep. 2015, pp. 1723–1728.
[10] D. Öhmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Diversity trade-offs and joint coding schemes for highly reliable wireless transmissions," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2016, pp. 1–6.
[11] *Study on Latency Reduction Techniques for LTE*, document 3GPP TR 36.881 v14.0.0, Jun. 2016.
[12] G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen, and P. Mogensen, "On the impact of multi-user traffic dynamics on low latency communications," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2016, pp. 204–208.
[13] B. Soret, K. I. Pedersen, M. C. Aguayo-Torres, and P. Mogensen, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 1391–1396.
[14] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jäntti, "Link adaptation design for ultra-reliable communications," in *Proc. IEEE Int. Conf. Commun.*, May 2016, pp. 1–5.
[15] 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)*, May 2017, pp. 1005–1010.
[16] C. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design,"in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.
[17] K. I. Pedersen, G. Pocovi, J. Steiner, and S. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Sep. 2017, pp. 1–6.
[18] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127–134, Dec. 2005.
[19] T. E. Kolding, "QoS-aware proportional fair packet scheduling with required activity detection," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2006, pp. 1–5.
[20] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 74–81, Apr. 2009.
[21] W. Nam, D. Bai, J. Lee, and I. Kang, "Advanced interference management for 5G cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 52–60, May 2014.
[22] B. Soret and K. I. Pedersen, "On-demand power boost and cell muting for high reliability and low latency in 5G," in *Proc. IEEE Veh. Technol. Conf.*, Jun. 2017, pp. 1–5.
[23] B. Soret, G. Pocovi, K. I. Pedersen, and P. Mogensen, "Increasing reliability by means of root cause aware HARQ and interference coordination," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2015, pp. 1–5.
[24] N. Brahmi, O. N. C. Yilmaz, K. W. Helmersson, S. A. Ashraf, and J. Torsner, "Deployment strategies for ultra-reliable and low-latency communication in factory automation," in *Proc. IEEE Globecom Workshops*, Dec. 2015, pp. 1–6.
[25] 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," in *Proc. IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2016, pp. 1–8.
[26] C. H. Yu, A. Hellsten, and O. Tırkkonen, "Rate adaptation of AMC/HARQ systems with CQI errors," in *Proc. IEEE Veh. Technol. Conf.*, May 2010, pp. 1–5.
[27] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.
[28] 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.
[29] D. Laselva, F. Capozzi, F. Frederiksen, K. I. Pedersen, J. Wigard, and I. Z. Kovacs, "On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2009, pp. 1–5.
[30] H. Holma and A. Toskala, *LTE Advanced: 3GPP Solution for IMT-Advanced*. New York, NY, USA: Wiley, 2011.
[31] K. I. Pedersen, T. E. Kolding, F. Frederiksen, I. Z. Kovacs, D. Laselva, and P. E. Mogensen, "An overview of downlink radio resource management for UTRAN long-term evolution," *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 86–93, Jul. 2009.
[32] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Veh. Technol. Conf.*, May 2000, pp. 1854–1858.
[33] B. Clerckx, H. Lee, Y.-J. Hong, and G. Kim, "A practical cooperative multicell MIMO-OFDMA network based on rank coordination," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1481–1491, Apr. 2013.
[34] V. Fernández-López, K. I. Pedersen, and B. Soret, "Interference characterization and mitigation benefit analysis for LTE-A macro and small cell deployments," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 1–12, Apr. 2015.
[35] M. Lampinen, F. Del Carpio, T. Kuosmanen, T. Koivisto, and M. Enescu, "System-level modeling and evaluation of interference suppression receivers in LTE system," in *Proc. IEEE Veh. Technol. Conf.*, May 2012, pp. 1–5.
[36] K. Brueninghaus *et al.*, "Link performance models for system level simulations of broadband radio access systems," in *Proc. IEEE Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2005, pp. 2306–2311.
[37] D. Chase, "Code combining—A maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. TCOMM-33, no. 5, pp. 385–393, May 1985.
[38] G. Pocovi, K. I. Pedersen, and B. Soret, "On the impact of precoding errors on ultra-reliable communications," in *Multiple Access Communications* (Lecture Notes in Computer Science), vol. 10121. Cham, Switzerland: Springer, 2016, pp. 45–54.
[39] L. D. Brown, T. T. Cai, and A. Dasgupta, "Confidence intervals for a binomial proportion and asymptotic expansions," *Annu. Statist.*, vol. 30, no. 1, pp. 160–201, 2002.

**GUILLERMO POCOVI** received the M.Sc. degree in telecommunications engineering from Universitat Politècnica de Catalunya in 2014, and the Ph.D. degree from Aalborg University, Denmark, in 2017. He currently holds industrial post-doctoral position at Nokia Bell Labs, Aalborg, partly sponsored by the Danish Innovation Fund. His research interests are mainly related to ultra-reliable and low-latency communications for current wireless networks and upcoming 5G New Radio.

**KLAUS I. PEDERSEN** 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 currently leading the Nokia Bell Labs Research Team, Aalborg, and a part-time Professor with in the Wireless Communication Networks Section, Aalborg University. He is currently involved in 5G New Radio, including radio resource management aspects, and the continued long-term evolution and its future development, with a special emphasis on mechanisms that offer improved end-to-end (E2E) performance delivery. He is currently part of the EU-funded research project ONE5G that focus on E2E-aware optimizations and advancements for the network edge of 5G New Radio. He has authored or co-authored approximately 160 peer-reviewed publications on a wide range of topics. He has invented several patents.

**PREBEN MOGENSEN** received the M.Sc. and Ph.D. degrees from Aalborg University in 1988 and 1996, respectively. He is currently a Principal Scientist with Nokia Bell Labs, Aalborg, Denmark, and a Bell Labs Fellow. He is also a Professor with Aalborg University and the Head of the Wireless Communication Networks Section, Aalborg University. He is currently involved in research and standardization for vertical use cases for LTE and 5G, including LPWA IoT, URLLC, I.4.0, V2X, UAV, and train communication. He has published over 400 papers within wireless communication and he has over 13 000 Google Scholar citations.