Preemptive Scheduling of Latency Critical Traffic and its Impact on Mobile Broadband Performance

Pedersen, Klaus I.; Gerardino, Guillermo Andrés Pocovi; Steiner, Jens

Published in:
2018 IEEE 87th Vehicular Technology Conference, VTC Spring 2018 - Proceedings

DOI (link to publication from Publisher):
10.1109/VTCSpring.2018.8417637

Publication date:
2018

Document Version
Early version, also known as pre-print

Preemptive Scheduling of Latency Critical Traffic and its Impact on Mobile Broadband Performance

Klaus I. Pedersen[1,2], Guillermo Pocovi[1,2], Jens Steiner[1]
Nokia – Bell Labs[1], Aalborg University[2]

Abstract— **In this paper, we present an exhaustive system level analysis of using preemptive scheduling for latency critical traffic in coexistence with mobile broadband for the 3GPP 5G New Radio. Enhanced recovery and HARQ retransmission mechanisms exploiting base station a priori knowledge of punctured radio resources from using preemptive scheduling are proposed. It is demonstrated that a scheme with HARQ multi-bit feedback and selective retransmission of punctured resources is an attractive solution. Furthermore, the performance sensitivity from using either fully interleaved or frequency-first code block layouts is assessed. The impact on the mobile broadband performance is evaluated at TCP-level, studying both the penalty on throughput and smoothened TCP round trip time to assess how preemptive scheduling affects the end-to-end performance of other traffic.**

I. INTRODUCTION

Research on the 5G New Radio (NR) is rapidly progressing with 3GPP about to release the first 5G specifications [1]. The ambitions for 5G NR are high, aiming for enhanced support for multiplexing of diverse services such as enhanced mobile broadband (eMBB) and low latency communication (LLC) with ultra-reliability constraints [2]-[4]. Simultaneously fulfilling the requirements for a mixture of users with such diverse QoS requirements is a challenging task [5], where the next generation base station (called gNB) scheduler plays an important role. So far, various 5G scheduling studies have been presented in the open literature, taking advantage of the flexible frame structure that comes with the 5G NR. Including cases with dynamic scheduling with variable transmission time intervals (TTI) [6]-[7], optimized scheduling for LLC with ultra-reliability constraints [8]-[9], and lately also punctured scheduling for enhanced downlink multiplexing of eMBB and LLC traffic with different TTI sizes [10]. The punctured scheduling scheme has similarities to preemptive scheduling principles as studied extensively for computer networks to accommodate real-time services [11]. This principle is now adopted by 3GPP for the 5G NR specifications under the name of preemptive scheduling. In short, preemptive scheduling allows to instantly schedule a latency critical transmission with a short TTI that fully or partly overwrites an ongoing eMBB transmission that uses a longer TTI size.

In this paper we further analyze the performance of preemptive scheduling, and propose related radio resource management policies to minimize the impact on eMBB users that experience preemption, where part of its transmission is overwritten, i.e. corrupted. We build on the study in [10] with the following additional enhancements: optimized hybrid automatic repeat request (HARQ) schemes, where only the damaged part of the preempted eMBB transmission(s) is retransmitted at first, and multi-bit HARQ feedback in the form of code-block-based ACK/NACK. It is analyzed how the performance of such solutions depend on the code-block (CB) layout of the eMBB transport block (TB), particularly the difference between fully interleaved random CB mapping (as assumed in [10]) and the so-called frequency-first mapping. In order to quantify the End-to-End (E2E) impact on the eMBB users, we model the effects of the transmission control protocol (TCP) [12], including the related flow control mechanisms. That means including the well-known slow start TCP procedure, which allows us to quantify if the preemptive scheduling triggers additional eMBB performance penalties as compared to earlier studies assuming simpler full buffer eMBB traffic models (i.e. without TCP effects). The proposed methods are evaluated in a dynamic multi-user, multi-cell setting with high degree of realism. Due to the complexity of the 5G NR system and the addressed problems, we rely on advanced system-level simulations for results generation. Those simulations are based on commonly accepted mathematical models, calibrated against the 3GPP 5G NR assumptions [1]-[2], making sure that statistical reliable results are generated.

The rest of the paper is organized as follows: Section II introduces the system model and presents the problem formulation and related objectives. The preemptive scheduling method is outlined in Section III, and the corresponding RRM considerations in Section IV. The performance analysis appears in Section V, followed by concluding remarks in Section VI.

II. SETTING THE SCENE

A. Basic system model

We adopt the 5G NR assumptions as outlined in [1]-[2], focusing primarily on the downlink performance. Users are dynamically multiplexed on a shared channel, using orthogonal frequency division multiple access (OFDMA). We assume a configuration with 15 kHz subcarrier spacing. LLC UEs are scheduled with short TTI of only 2 OFDM symbols, corresponding to a mini-slot of 0.143 ms. eMBB traffic is primarily scheduled with longer TTI sizes of 14 OFDM symbols (1 ms duration), equivalent to one slot (but could also be scheduled with shorter TTI sizes). In the frequency domain, users are multiplexed on a physical resource block (PRB) resolution of 12 subcarriers. Users are dynamically scheduled, using a user-centric downlink control channel for transmitting the scheduling grant [13]. This includes informing the users on which resources they are scheduled, which modulation and coding scheme (MCS) is used, etc. Asynchronous HARQ with Chase combining is assumed. The system is assumed to carry best effort eMBB traffic download, as well as sporadic LLC traffic. The latter is modeled as bursts of small payload size of B bits that arrive for each LLC user in the downlink direction following a uniform Poisson arrival point process with arrival

rate λ. Thus, the offered LLC traffic load per cell equals $N \cdot B \cdot \lambda$, where N is the average number of LLC users per cell.

B. Problem formulation and KPIs

The objective is to serve the eMBB users with high average data rates (i.e. maximizing their spectral efficiency), while serving the LLC users per their low latency requirement with ultra-high reliability. Hence, for the eMBB users the primary key performance indicator (KPI) is the TCP throughput, but also the round trip time (RTT) of TCP packets is monitored, defined as the time from the server generates the TCP packet until the corresponding Ack is received. In line with the definition in RFC6298, the smoothened TCP RTT (sRTT) is considered. The LLC traffic takes priority over the best effort eMBB data flows, and needs to be immediately scheduled when it arrives at the gNB. The primary KPI for the LLC traffic is the latency from the time the payload arrives at the gNB until it is successfully received at the UE, i.e. the one-way downlink RAN performance for this service type.

III. Preemptive Scheduling Principle

A. Basic principle

The basic principle of the preemptive scheduling solution is illustrated in Fig. 1, showing how multiple users are time-frequency domain multiplexed on the downlink shared channel. The default operation here is that the eMBB users are scheduled with a TTI size of one or multiple slots. During the transmission time of the TBs for the eMBB UEs, the shared channel is in principle monopolized on those resources. However, it may happen that LLC data for another UE arrives at the gNB while the scheduled transmission towards the eMBB UEs is ongoing. To avoid waiting for the completion of the TB transmission to the eMBB UEs, preemptive scheduling allows the gNB to immediately transmit the LLC data by puncturing (i.e. overwriting) part of the ongoing eMBB transmission. This is accomplished by scheduling the LLC payload with a short TTI of one mini-slot. The advantage is that the latency of the LLC data is minimized. However, as some of the resources for the eMBB transmission(s) are corrupted, it essentially results in an error floor, where the performance in terms of block error probability (BLEP) vs SINR for the UE saturates [14].

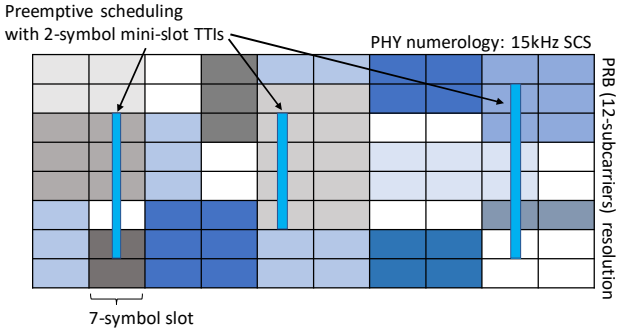

Fig. 1: Basic principle of downlink preemptive scheduling.

The impact on the eMBB UE(s) performance from the preemptive scheduling depends on multiple factors, including how the information bits for the eMBB TB have been encoded, interleaved, and mapped to the physical layer resources [14]. In line with the NR assumptions [1], we assume that an eMBB transmission consists of CBs. The maximum CB size equals Z=6144 bits, and the number of CBs is denoted by C [15]. For the sake of simplicity, we assume that the CBs have equal size. In line with the study in [10], the *baseline CB layout* corresponds to fully time-frequency interleaving over the assigned resources for the TB. Secondly, we also study the case with the so-called *frequency-first CB mapping* scheme, where individual CBs are spread over the allocated frequencies, but condensed in time-domain. The two considered CB layouts are illustrated in Fig. 2, where the colors represent different CBs within the TB. Under conditions where the eMBB transmission is not subject to preemption, and the SINR is similar for all CBs, the error probability for the TB can be expressed as $P(\varepsilon_{TB}) = 1 - (1-P(\varepsilon_{CB}))^C$, where $P(\varepsilon_{CB})$ is the CB BLEP. However, for cases where the TB is subject to partial preemption, the situation changes as follows: for the fully interleaved case, the CB BLEP will be equally degraded for all CBs, while for the frequency first case the CBs will be unequally affected. If the preemption e.g. happens to only overwrite CB#3 and CB#2 (example in Fig. 2), then the BLEP for those two CBs will approach 100%, while the BLEP for CBs #1, #4, and #5 will remain unaffected of the preemptive scheduling. Therefore, when a TB transmission is subject to partial preemption, the usage of fully interleaved vs frequency-first CB layout presents a tradeoff between a smaller equal degradation of the per CB BLEP vs affecting only a few of the CBs with much stronger impact (i.e. BLEP approaching 100%).

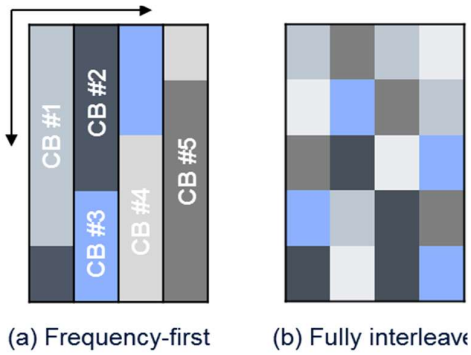

Fig. 2: Basic illustration of the two considered code black (CB) layouts: (a) frequency-first and (b) fully interleaved CB layout.

B. Recovery mechanisms and enhanced HARQ operation

The decoding probability (i.e. 1-BLEP) of the preempted eMBB TB transmission depends on whether the UE is aware of this happening. In line with [10],[14], the performance is improved if the eMBB UE is aware of the preemption. The former is facilitated by sending a *preemption indication (PI)* to affected eMBBs, such that they know which resources of the

Table 1: Summary of considered recovery and HARQ schemes.

HARQ scheme	Description	UE feedback
#1	Baseline: upon reception of NACK, the full TB is retransmitted.	Single-bit ACK/NACK
#2	Partial retransmission: if the TB was subject to preemption, only the damaged part of the TB is retransmitted. Only if a second NACK is received for the same HARQ process is the full TB retransmitted again.	Single-bit ACK/NACK
#3	Only the CBs with NACK feedback are retransmitted. This is done independently on whether the TB was subject to preemptive scheduling, or not.	CB-based ACK/NACK (multi-bit feedback)
#4	If the first TB transmission was subject to preemption, only the damaged parts of the CBs having received a corresponding NACK for is retransmitted. If the full TB is not correctly decoded after the second HARQ transmission, the full TB is retransmitted again. If the first TB transmission was not subject to preemption, HARQ scheme #3 is applied.	CB-based ACK/NACK (multi-bit feedback)

TB transmission have been corrupted. The eMBB UE(s) benefit from preemptive indication information by disregarding the corrupted resources of the transmission in its decoding process, including potentially performing HARQ soft combining, thereby improving the performance. Throughout this study, we assume that the eMBB users experiencing preemption are informed. We refer to [10],[14], for performance results for cases where eMBB users are unaware of the preemption event.

In case the UE fails to correctly decode the eMBB transmission, a HARQ retransmission is sent by the gNB. We consider four different retransmission strategies as follows: As a baseline, we consider the cases with a single-bit ACK/NACK feedback from the UE, where the complete TB is retransmitted, consuming the same amount of transmission resources as the original transmission on the same HARQ process. Secondly, we consider an enhanced HARQ retransmission strategy for eMBB transmissions that have suffered from preemption. We label this option as *partial HARQ retransmission*, as the gNB only retransmit the damaged part of preempted TB upon reception of the NACK. If the UE again feeds back a NACK for the same HARQ process, the full TB will be retransmitted (i.e. third transmission).

We furthermore consider cases where multi-bit ACK/NACK feedback is provided by the UE in the form of separate ACK/NACK per CB. Given the CB-based ACK//NACK feedback, the third HARQ scheme simply retransmits the entire CBs for which the gNB has received a NACK. Finally, the fourth scheme, assumes that only the potentially damaged parts of NACK'ed CBs that have been subject to preemption is retransmitted. The four considered HARQ schemes are summarized in Table 1. As will be shown in the subsequent sections, the advantage of HARQ schemes 2-4 is that the resource usage for HARQ retransmissions is more efficiently optimized as compared to the baseline (i.e. HARQ scheme #1).

IV. RRM Algorithms

A. Scheduling decisions

For scheduling of the eMBB traffic we assume time-frequency domain radio channel aware Proportional Fair (PF) scheduling, based on periodical frequency selective CQI feedback. The PF scheduling metric $M_{u,p}$ is:

$$M_{u,p}[n] = \frac{r_{u,p}[n]}{R_u[n]}, \quad (1)$$

where $r_{u,p}$ is an estimate of the instantaneous supported data rate of user u in the p-th PRB, R_u is its average delivered throughput in the past, and n is the discrete time index for the scheduling interval. eMBB users are scheduled with a TTI size of 1 ms. Pending eMBB HARQ retransmissions are prioritized over new eMBB transmissions as also assumed in [15]. By default, the eMBB traffic is scheduled on all available radio resources, assuming there is enough offered eMBB traffic.

When LLC traffic arrives at the gNB, the scheduler aims at immediately scheduling such traffic with a short TTI size of 0.143 ms (corresponding to 2 OFDM symbols). If there are free (unused) radio resources, the LLC traffic is scheduled on those resources. If not, the LLC traffic is scheduled on radio resources currently allocated to eMBB transmissions, using preemptive scheduling according to the PF rule. For alternative eMBB-aware scheduling algorithms with preemptive allocations of LLC traffic, we refer to the recent study in [10].

B. Service-specific link adaptation

Dynamic link adaptation (LA) is assumed for both the eMBB and LLC users by setting the MCS for each transmission based on the users reported CQI. The MCS for the eMBB users is adjusted to reach an average block error rate (BLER) target of 10%. This is achieved by using the well-known outer loop link adaptation (OLLA) algorithm, where the received CQI values are offset by certain factor (a.k.a. the OLLA offset) calculated in accordance to the received HARQ Ack/Nacks from past transmissions [16]. In line with [10], the 10% BLER target for the eMBB users is only maintained for transmissions that are not subject to preemption, while eMBB transmissions suffering from preemption will experience a higher BLER.

The LA for the LLC transmissions is conducted to have a BLER target of only 1% to have lower latency. The LA for the LLC users is also conducted based on the users CQI, using standard OLLA to reach the 1% BLER target.

V. Performance Analysis

A. Methodology and assumptions

Extensive dynamic system-level simulations are conducted, following the 5G NR methodology in 3GPP [1], [3], assuming a macro-cellular multi-cell scenario. The default simulation

assumptions are summarized in Table 1. Five eMBB users are present per cell, spatially uniformly distributed. Users connect to the cell corresponding to the highest received power. There is a finite downlink payload of 50 kB for each eMBB user, using the TCP. The TCP implementation follows the Cubic model. When a TCP packet (with maximum segment size – MSS - of 1500B) is generated at the traffic source, it is subject to a core network (CN) latency of 2 ms before arriving at the base station. The corresponding TCP acknowledgement (Ack) from traffic sink (UE) in the uplink is transmitted with the same TTI size as in the downlink. Conveying the TCP Ack from the base station to the traffic source is again subject to the CN latency. Whenever an eMBB user completes its file download, the user is removed from the system, and a new eMBB user is generated at a random location in the cell. Thereby ensuring that the network contains a constant number of five eMBB users per cell throughout the simulation.

Table 1: Summary of default simulation assumptions.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector base stations with 500 meters inter-site distance. 21 cells.
Carrier	10 MHz carrier bandwidth at 2 GHz (FDD)
PHY numerology	15 kHz subcarrier spacing configuration [1].
TTI sizes	0.143 ms for LLC (2-symbol mini-slot). 1 ms for eMBB (two slots of 7-symbols).
MIMO	Single-user 2x2 closed loop MIMO and UE MMSE-IRC receiver.
CSI	Periodic CSI every 5 ms, with 2 ms latency, containing CQI, and PMI.
Data channel modulation and coding	QPSK to 64QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS selection. 1% initial BLER target for LLC 10% initial BLER target for eMBB
HARQ	Asynchronous HARQ with Chase Combining. The HARQ RTT equals minimum 4 TTIs.
Traffic model	5 finite buffer eMBB users per cell with 50 kbyte file download over TCP. 10 LLC users per cell with Poisson arrival of B=50 bytes data bursts.
Scheduling	Proportional fair scheduling of eMBB. LLC traffic always take priority over eMBB, using preemptive scheduling if no free resources.
Transport layer *(only relevant for eMBB)*	TCP Cubic model, RFC 5681 TCP MSS: 1500B Initial TCP Window: 3xMSS SSThreshold: 45xMSS=67.5kB One-way Core Network (CN) delay: 2 ms
Link-to-system (L2S) mapping	Based on the mean mutual information per coded bit (MMIB) mapping methodology.

For the LLC traffic, we apply a bursty stochastic model, with 50-byte packets generated according to a homogeneous Poisson arrival process. Different levels of offered LLC traffic load per cell are considered. The LLC payloads arrives directly at the gNBs, and hence effects of CN delays and TCP flow control mechanisms are not considered from this traffic type. The former corresponds to the case where the LLC payloads are unstructured datagrams as is the typical case for machine-type-communication.

Whenever a user is scheduled, the SINR at the receiver is calculated for each subcarrier symbol, assuming a minimum mean square error with interference rejection combining (MMSE-IRC) receiver at the terminal. Inspired by the model in [17]-[18], the SINR values are mapped to the mutual information domain, taking the applied modulation scheme into account. The mean mutual information per coded bit (MMIB) is calculated as the arithmetic mean of the values for the sub-carrier symbols of the transmission [18]. Given the MMIB and the used modulation and coding rate of the transmission, the error probability of a CB is determined from look-up tables that are obtained from extensive link level simulations. For transmissions consisting of more than one CB, the CB BLEP is calculated per CB, and only if all CBs in the TB are correctly received does the UE send a single-bit ACK.

The effect of an eMBB transmission that is preempted is captured as follows: The corrupted sub-carrier symbols containing no useful information for the receiver are modelled as information-less. UEs are assumed to be aware of the preemption. Therefore, the receiver discards the punctured parts of the physical resources prior to the decoding. Hence, the MMIB for such users is calculated only as the mean from transmission resources that were not punctured and the effective coding rate of the transmission is increased accordingly. See [10] for cases where the UE is unaware of the preemption.

B. Performance results

We first present average TCP-level throughput results for the eMBB users versus the offered LLC traffic per cell. The LLC traffic is increased up to an average offered load of 2 Mbps/cell, which corresponds to approximately using 12% of the radio resources on average for such traffic. Fig. 3 shows the performance for the case with fully interleaved CB layout, while Fig. 4 shows similar results for the frequency first CB layout. As expected, the eMBB throughput decreases as the LLC load is increased. This is a consequence of prioritizing LLC traffic by using preemptive scheduling to overwrite part of the eMBB transmissions. For the fully-interleaved CB layout, the HARQ schemes with partial retransmission (scheme #2 and #4) provide the best performance, as fewer resources are used for retransmissions, thereby unleashing more resources for transmissions to other eMBB users. Multi-bit HARQ feedback on its own (scheme #3) does not provide significant gain over the baseline (scheme #1), since the CBs tend to experience similar BLEP with high correlation.

In contrast, HARQ scheme #3 provides a significant gain for cases with frequency-first CB layout (Fig. 4). This is because the preemption typically affects only one or very few CBs, hence the retransmission size can be significantly reduced if only damaged CBs are retransmitted. For frequency first CB layout, HARQ scheme #4 provides the best throughput performance: up to 30% throughput gain as compared to the

baseline HARQ scheme #1, at 2 Mbps LLC offered load. When comparing the results in Fig. 3 and Fig. 4, the fully-interleaved CB layout provides the best performance for HARQ schemes #1, #2, and #4. This is because having modest equal impact on all CBs reduces the TB error probability as compared to the case where only a few CBs are damaged. The actual throughput gain of fully interleaved over frequency first CB layout varies depending on the HARQ scheme. For instance, a relevant 10% throughput improvement is obtained for HARQ scheme #1 at 2 Mbps LLC offered load; whereas the gain is only 1% for HARQ scheme #4. For HARQ scheme #3, the frequency-first CB layout results in the best performance at 2 Mbps URLLC load. This is because this scheme relies solely on the multi-bit CB-level ACK/NACK feedback from the UE, and hence less radio resources are spent on HARQ retransmissions if the impact from the preemption is limited to few CBs only. It is worth mentioning that frequency-first CB layout offers additional benefits in terms of pipeline processing, as the UE can decode the first CBs even before having received the full TB.

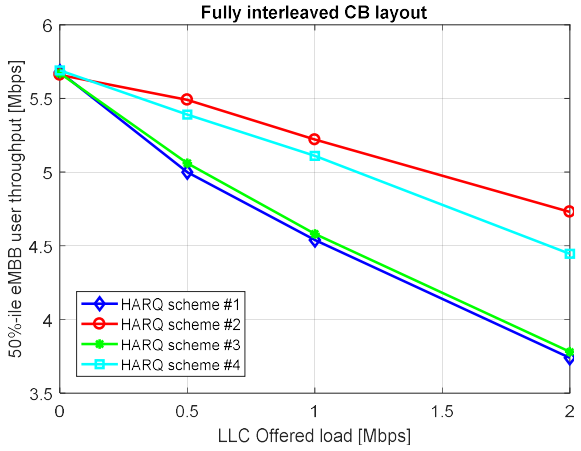

Fig. 3: Aggregated TCP-level cell throughput for the eMBB users, assuming the fully interleaved CB layout.

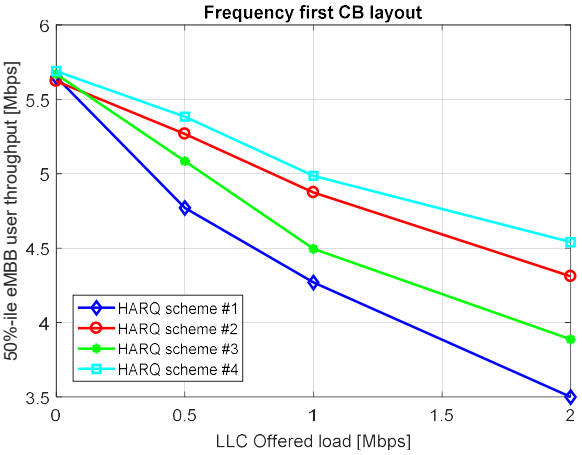

Fig. 4: Aggregated TCP-level cell throughput for the eMBB users, assuming the frequency first CB layout.

Fig. 5 shows the complementary cumulative distribution function (ccdf) of the latency performance for the LLC traffic. Three curves are shown: low (500 kbps/cell LLC traffic), medium (1 Mbps/cell LLC traffic) and high (2 Mbps/cell LLC traffic). It is observed that the stringent URLLC target of 1 ms and 99.999% reliability is fulfilled for the three cases. In line with link adaption target of 1% BLEP for first LLC transmissions, 99% of the LLC transmissions are successfully received on the initial transmission attempt, with a latency as low as 0.3 ms.

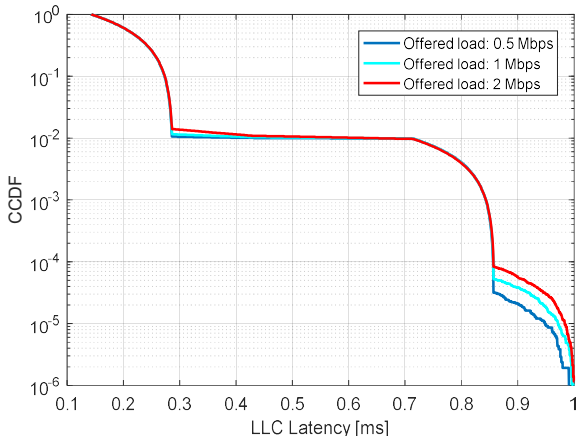

Fig. 5: Complementary cumulative distribution function of the latency performance for the LLC traffic.

We next take a closer look at the impact from the preemptive scheduling on the eMBB performance at TCP level. Fig. 6 summarizes the relative degradation of TCP-layer throughput and sRTT when increasing the URLLC load from 0 Mbps to 2 Mbps. We show statistics for the median, 5% worst and 5% best eMBB users, assuming frequency first CB layout. In line with the findings in [10], the lowest impact of preemption is observed for the 5% worst eMBB users (5% and 95% percentile of the throughput and sRTT distribution, respectively). This is because such users are typically scheduled with a low MCS order, and hence are more tolerant to preemption (see [10] for more details). On the other hand, the largest performance degradation is observed for the 5% best eMBB users (95% and 5% percentile of the eMBB throughput and sRTT distribution), as allocations with high MCS order are more sensitive to preemption. Those 5% eMBB users that have the best radio conditions, are also those that are relatively influenced the most by the slow start TCP procedure, and most sensitive to rare events of triggering a slow start TCP; e.g. due to being preempted by LLC traffic.

Fig. 6 also shows the benefit of employing enhanced HARQ schemes. HARQ scheme #4 generally offers better performance in terms of throughput and sRTT, as compared to HARQ scheme #1. At the median (50%-fractile), the degradation of TCP layer throughput, and increased sRRT, is only on the order of 20% from carrying an additional LLC traffic of 2 Mbps/cell, where every LLC payload of 50 bytes is timely delivered within 1 ms with a reliability level of 99.999%. Thus, this is in line

with theory [5]: With no latency constraints, the effective capacity equals the Shannon capacity, while it decreases asymptotically as stricter latency constraints are enforced.

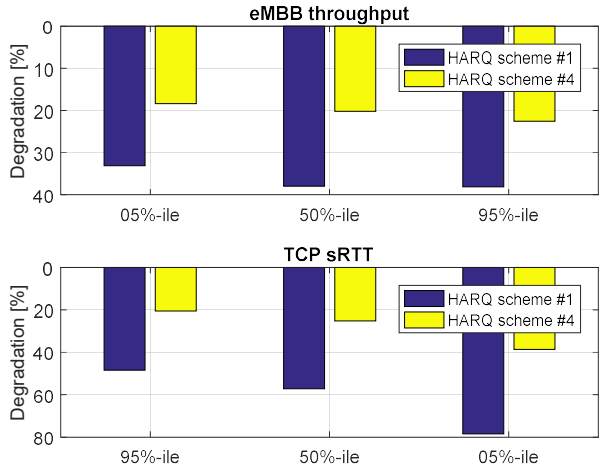

Fig. 5: Summary of the impact on eMBB TCP performance from preemptive scheduling of LLC traffic.

VI. Concluding Remarks

In this study we have presented a detailed system-level performance evaluation of preemptive scheduling for LLC traffic. The presented performance results are encouraging, showing good performance of this technique for multiplexing LLC and eMBB traffic without the need of pre-reserving radio resources for sporadically arriving LLC data bursts. Various HARQ retransmission mechanisms have been proposed to efficiently recover eMBB transmission that suffer from preemption. It has been shown that HARQ schemes with multi-bit feedback are the best, exploiting base station's a priori knowledge to only retransmit the punctured parts of eMBB transmissions. For an LLC offered load of 2 Mbps/cell, such enhanced retransmission schemes provide up to 30% higher eMBB throughput as compared to the baseline HARQ scheme where the entire transport block is retransmitted. These techniques are especially relevant for frequency-first CB layouts, where the effect of preemption is typically limited to only a few CBs.

Acknowledgements

Part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project. Secondly, the work is also partly funded by the Innovation Fund Denmark (IFD) under File No. 7039-00009B.

References

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN Overall Description; Stage-2", June 2017.

[2] 3GPP Technical Report 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies", March 2016.

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

[4] E. Dahlman, et.al., "5G Wireless Access: Requirements and Realization", IEEE Communications Magazine - Communications Standards Supplement, December 2014.

[5] B. Soret, et.al., "Fundamental Tradeoffs among Reliability, Latency and Throughput in Cellular Networks", IEEE Proc. GLOBECOM, December 2014.

[6] Q. Liao, P. Baracca, D. Lopez-Perez, L.G. Giordano, "Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems", in IEEE Proc. Globecom, December 2016.

[7] K.I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, S.R. Khosravirad, "System Level Analysis of Dynamic User-Centric Scheduling for a Flexible 5G Design", in IEEE Proc. Globecom, December 2016.

[8] G. Pocovi, B. Soret, K.I. Pedersen, P.E. Mogensen, "MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks", in IEEE Proc ICC (workshop), June 2017.

[9] G. Pocovi, K.I. Pedersen, B. Soret, M. Lauridsen, P.E. Mogensen, "On the Impact of Multi-User Traffic Dynamics on Low Latency Communications", in Proc. International Symposium on Wireless Communication Systems (ISWCS), September 2016.

[10] K.I. Pedersen, G. Pocovi, J. Steiner, S. Khosravirad, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband", in IEEE Proc. VTC-fall, September 2017.

[11] G.C. Buttazzo, M. Bertogna, G. Yao, "Limited Preemptive Scheduling for Real-Time Systems: A Survey", in IEEE Trans. on Industrial Informatics, vol. 9, no. 1, pp. 3-15, Feb. 2013.

[12] A.S. Tanenbaum, "Computer networks", fifth edition, Prentice Hall, 2011.

[13] K.I. Pedersen, et.al., "A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases", in IEEE Communications Magazine, pp. 53-59, March 2016.

[14] Technical contribution to 3GPP, Document R1-1700374, "Downlink Multiplexing of eMBB and URLLC Transmissions", Intel Corporation, January 2017.

[15] 3GPP TS 36.212, "Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding", January 2017.

[16] H. Holma and A. Toskala (Editors), "LTE-Advanced: 3GPP solution for IMT-Advanced", John Wiley & Sons, 2012.

[17] A. Pokhariyal, et.al., "HARQ Aware Frequency Domain Packet Scheduling with Different Degrees of Fairness for UTRAN Long Term Evolution", in IEEE Vehicular Technology Conference (VTC-Spring), May 2007.

[18] K. Brueninghaus, et.al, "Link performance models for system level simulations of broadband radio access systems," in IEEE Proc. Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 2306-2311, Sept. 2005.

[19] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, "IEEE 802.16m Evaluation Methodology Document (EMD)", in IEEE 802.16 Broadband Wireless Access Working Group, Tech. Rep. IEEE 802.16m-08/004r2, http://ieee802.org/16/tgm/docs/80216m-08_004r2.pdf, July 2008.