On the Impact of Multi-User Traffic Dynamics on Low Latency Communications

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

Published in:
Wireless Communication Systems (ISWCS), 2016 International Symposium on

DOI (link to publication from Publisher):
10.1109/ISWCS.2016.7600901

Publication date:
2016

Document Version
Accepted author manuscript, peer reviewed version

On the Impact of Multi-User Traffic Dynamics on Low Latency Communications

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

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

***Abstract*—In this paper we study the downlink latency performance in a multi-user cellular network. We use a flexible 5G radio frame structure, where the TTI size is configurable on a per-user basis according to their specific service requirements. Results show that at low system loads using a short TTI (e.g. 0.25 ms) is an attractive solution to achieve low latency communications (LLC). The main benefits come from the low transmission delay required to transmit the payloads. However, as the load increases, longer TTI configurations with lower relative control overhead (and therefore higher spectral efficiency) provide better performance as these better cope with the non-negligible queuing delay. The presented results allow to conclude that support for scheduling with different TTI sizes is important for LLC and should be included in the future 5G.**

I. Introduction

Fifth generation (5G) cellular technologies are expected to bring support for a wide range of use cases [1]-[3]. 5G is foreseen not only to cope with the continuously increasing mobile broadband (MBB) traffic demands, but also to enable novel communication paradigms such as ultra-reliable low-latency communications (URLLC) [2]-[4].

The downlink latency performance in a multi-user cellular network is the focus of this paper. Achieving low latency communication (LLC) is very challenging as it requires the optimization of the multiple components that contribute to the latency budget [5]. The queuing delay at base station nodes is a particularly important component. This is a function of the offered load, traffic dynamics, scheduling strategy and also aspects related to the air interface, e.g. frame structure and transmission time interval (TTI). Examples of studies investigating the queuing delay (and related system aspects) include the work in [6], where the tail distribution of the delay is estimated with different scheduling strategies over a time-slotted fading channel. In the context of cellular networks, the work in [7] analyses the delay performance of various multiple-access schemes with multiple priority classes. In [8], a discrete queuing model is applied to study the downlink throughput and delay performance of a orthogonal frequency division multiple access (OFDMA)-based system. More recently, the work in [9] proposes a flexible frame structure for dynamic scheduling of users with different TTI sizes in accordance to each user requirements. Although short TTI (e.g. 0.25 ms) is beneficial to reduce the over-the-air transmission time, it has a cost in terms of higher signalling overhead and therefore lower spectral efficiency [10]. There is therefore a compromise between the benefits of having short TTI durations, and the experienced queuing delay as a result of the reduced spectral efficiency.

In this work we go a step forward and analyse the tradeoffs between queuing delay and TTI size on a system level. Our main focus is on the achievable latency under different TTI durations and system loads; but we also present relevant results about the spectral efficiency and throughput performance under the different system configurations. We build on the recent study in [9], that proposes dynamic adjustment of the TTI on a per-user basis. The evaluation methodology is dynamic system-level Monte Carlo simulations with bursty traffic, where we consider the effects of different radio channel conditions per user, and varying relative control overhead depending on the TTI size. Despite some simplifications at the physical layer such as error-free transmissions, our simulation framework allows us to draw initial conclusions on the impact of different elements on the total latency, and relevant tradeoffs between spectral efficiency and latency. In a nutshell, our results reveal that as the load increases, the system must gradually increase the TTI size (and consequently the spectral efficiency) in order to cope with the non-negligible queuing delay.

The rest of the paper is organized as follows: Section II describes the multiple elements accounting for the user latency. Section III presents an overview of the considered frame structure, including multiplexing of users and scheduling format considerations. Section IV explains the methodology and considered assumptions. Performance results are presented in Section V, followed by a discussion in Section VI. Finally, conclusions are summarized in Section VII.

II. Latency Composition and Related Definitions

We first describe the various sources that contribute to the downlink latency in a cellular system. A traffic source generates data that are transmitted to a traffic sink via the cellular system. First, the data from higher layers are received at the base station node and are stored in the transmission buffers. Some time is typically required at the base station to process the data and perform the scheduling decision. When the payload is ready to be scheduled, the system must wait to the beginning of the next TTI to transmit the data, assuming a time-slotted system. The data is placed in the radio frame and transmitted to the mobile terminal, where it is subject to a certain processing delay before it is successfully decoded and forwarded to the traffic sink at higher layers. The user-plane one-way latency L for a user scheduled in the downlink can therefore be expressed as [5],

$$L = d_{\mathrm{Q}} + d_{\mathrm{bsp}} + d_{\mathrm{FA}} + d_{\mathrm{Tx}} + d_{\mathrm{mtp}} \quad [\mathrm{s}], \tag{1}$$

(a) Queuing delay: Time it takes for the packet to arrive to the transmission buffer

(b) Frame alignment delay: Time from the moment the packet arrives to the transmission buffer to the beginning of next TTI

(c) Transmission delay: Number of TTIs needed to transmit the packet

Fig. 1: Sketch of distribution of the different delay components.

where d_{Q}, d_{FA} and d_{Tx} represent the queuing, frame alignment, and transmission delay, respectively; and d_{bsp} and d_{mtp} represent the processing delay at the base station and mobile terminal. Note that we refer to *delay* as the separate contribution of the various components, and *latency* to the sum of all components. Some of these components are described in Fig. 1. The queuing delay depends on the amount of users that are multiplexed on the same radio resources. Given the random behaviour of packet arrivals, even at relatively low load, there is a probability of experiencing queuing delay due to the instantaneous variation of the incoming traffic. The frame alignment delay depends on the frame structure and duplexing mode. For frequency division duplex (FDD) modes, such as considered in this work, the frame alignment delay is bounded between 0 and the TTI duration, depending on whether the packet reaches the buffer right before or after a TTI begins. The transmission of the payload takes at least one TTI but it can take multiple TTIs depending on the available resources, payload size, radio channel conditions, transmission errors and the respective retransmissions, etc. The processing delay at both base station and mobile terminal depends on their processing capabilities and is typically on the order of a few milliseconds in LTE for each downlink data payload [5]. Shorter processing delay is expected for 5G in order to allow support for lower latency [2].

III. Overview of 5G Flexible Frame Structure

The OFDMA-based frame structure presented in [9] is adopted. Users are flexibly multiplexed on a grid of orthogonal time-frequency tiles, as shown in Fig. 2. Each tile corresponds to the minimal resource allocation for a user, composed of one subframe in the time domain and a physical resource block (PRB) in the frequency domain. On each scheduling opportunity, an arbitrary number of tiles can be assigned to each user providing therefore high flexibility in terms of TTI length and bandwidth allocation. The control channel (CCH), marked as dark blue in Fig. 2, is accommodated within the resources assigned to each user (i.e. in-resource CCH). The CCH contains the scheduling grant indicating the specific time-frequency resource allocation for each user, among other relevant link adaptation parameters required to decode the data. The actual resource allocation is performed in accordance with the user-specific service requirements. Using a short TTI (e.g. 0.25 ms) allows to achieve low frame alignment delay and shorter transmission time, at the expense of large CCH overhead. In contrast, the use of long TTIs results in lower CCH overhead, among other benefits that increase the spectral efficiency of the system [9].

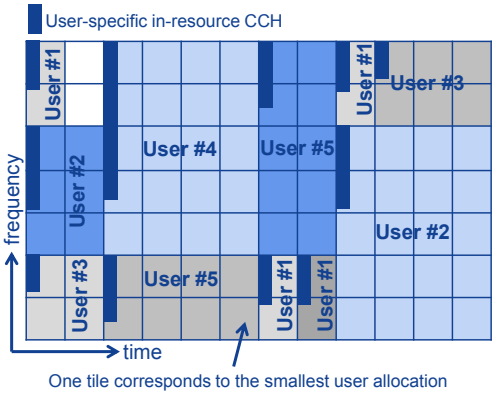

Fig. 2: User multiplexing example on 5G flexible frame structure.

A. Scheduling format and frame numerology

The CCH and data are multiplexed within the assigned resources per user. This user-specific approach allows to dynamically vary the coding rate of the CCH overhead in order to match the channel conditions of each user (note the difference in size of the user-specific CCH depicted in Fig. 2). Taking the LTE physical downlink control channel (PDCCH) link-performance as a reference, a minimum of 36 resource elements (REs) are required to transmit the CCH with a block-error rate (BLER) of 1% or less for users experiencing relatively good channel conditions [11]. One RE corresponds to one OFDM subcarrier symbol. Additional robustness is obtained by using higher aggregation levels (i.e. repetition encoding rate) of 2, 4, or 8. Table I summarizes the required number of REs for the CCH depending on the user-specific signal to interference and noise ratio (SINR) [11].

We adopt one of the physical layer numerology options proposed for 5G in [12]. It consists of 16 OFDM symbols per 1 ms, 17.143 kHz subcarrier spacing, and a PRB size of 12 subcarriers. We consider TTI durations of 0.25, 0.5, 1 and 2 ms (4, 8, 16 and 24 OFDM symbols, respectively). On every 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. This sets a constraint on the minimum allocatable resources to a user. As an example, for a TTI of 0.25 ms (4 OFDM symbols and 4x12 = 48 REs within one PRB), the minimal resource allocation to a user varies from 1, 2, 4, and 7 PRBs depending on its SINR value (see Table I), when including an additional 10% of reference symbol overhead [11]. For a more exhaustive study on 5G frame numerology options we refer to [12].

TABLE I: Control channel overhead [11]

SINR [dB]	In-resource CCH overhead
$(-\infty, -2.2)$	8x36 = 288 REs
$[-2.2, 0.2)$	4x36 = 144 REs
$[0.2, 4.2)$	2x36 = 72 REs
$[4.2, \infty)$	1x36 = 36 REs

IV. Simulation Framework

The performance evaluation is based on system-level simulations of a multi-user cellular system. Two types of traffic are simultaneously evaluated. (i) Bursty traffic with a finite payload of B bits per user with random arrivals that follow a Poisson process with arrival rate λ (the offered load is $\lambda \cdot B$). We refer to this traffic type as *LLC*. And (ii) full buffer traffic from a single user per cell with infinite payload of downlink data. The latter, referred to as *MBB*, allows us to analyse the impact of different system configurations on the throughput performance. The simulation procedure follows the diagram in Fig. 3. LLC users arriving to the system are assigned with a SINR randomly chosen from a given distribution. The SINR distribution is taken from a 3GPP regular macro cellular network with 500 m inter-site distance (ISD), where users are uniformly distributed. The SINR distribution captures the effects of distance-dependent attenuation, shadowing and full-load inter-cell interference according to [13]. This approach reproduces the different radio channel conditions depending on the location of the user in a cellular network. Explicit modelling of fast fading is not included. The transmitted data bits on a given time-frequency resource of size (t, bw) are given by,

$$N_{bits} = t \cdot bw \cdot \log_2(1 + SINR) \cdot \eta(t, bw, SINR) \quad [\text{bit}], \quad (2)$$

where t corresponds to the TTI duration, and bw is the bandwidth of the allocated resource composed of an integer number of PRBs. The transmission efficiency $\eta(t, bw, SINR)$ represents the relative CCH overhead of the (t, bw)-sized resource. This is calculated as the amount of REs used for the scheduling grant (given in Table I for different SINR values) plus an additional 10% for reference symbols, divided by the total amount of REs in the block of (t, bw) size. LLC users are scheduled with a first-come first-served (FCFS) policy with priority over MBB traffic. Since we are mainly interested in the tradeoffs between the queuing delay and the TTI size, we assume a fixed TTI size per simulation for both MBB and LLC traffic. After the payload of B bits is delivered, the call is terminated. Frequency multiplexing of users can occur for the cases where the transmission of a certain payload occupies less than the available resources in a TTI. Table II summarizes the default simulation assumptions.

Simulations are run with different offered loads and TTI durations, and relevant statistics are obtained for each type of traffic. The main performance indicators for MBB and LLC are, respectively, the downlink experienced throughput and the latency, as defined in Section II. The processing delay is assumed to be constant for each call, and is therefore not included in the simulations. The simulation time corresponds to at least 100.000 calls to ensure a reasonable confidence level for the considered performance measures.

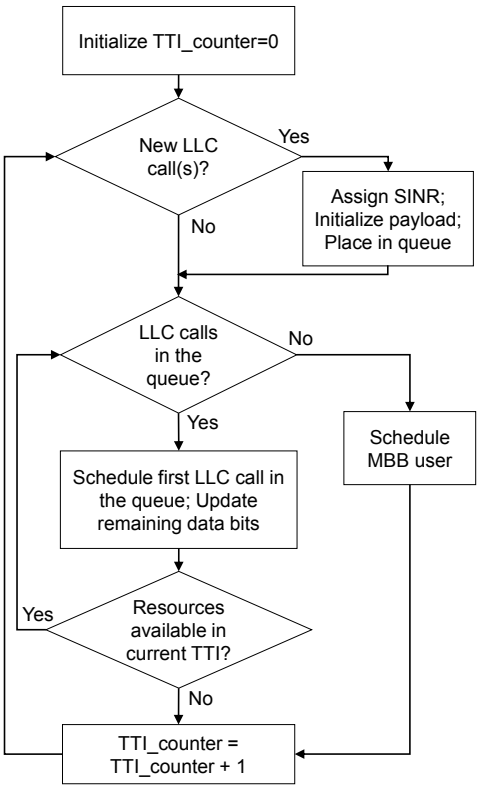

Fig. 3: Flow diagram of simulation methodology.

V. Results

We start by analysing the impact of different TTI sizes on the MBB throughput performance. Table III summarizes the MBB user throughput for different SINR and offered loads of LLC traffic. An offered load of 4 Mbps corresponds to an average LLC resource utilization of approximately 25%. As expected, the throughput decays when the load increases. It can also be observed how the throughput is affected by the TTI size. Shorter TTIs result in larger CCH overhead and in consequence lower spectral efficiency. For example, the gain from using a 2 ms TTI over a 0.25 ms TTI is, respectively, 17% and 20% for the lowest and highest offered load and a SINR of -3 dB. At +3 dB SINR, the gain from long TTI is reduced to 4% and 8% for 4 Mbps and 12 Mbps offered load, respectively. In general, the largest gains from using a long TTI size are obtained at low SINR. This is mainly due to the larger impact of the CCH overhead for users experiencing poor radio channel conditions.

Fig. 4 presents the latency at the 50% (median) and 99% percentile for different offered loads and TTI durations. At the median, it is shown that the achievable latency is not significantly impacted by the offered load. In this case, the dominant components of the latency budget are mainly the frame alignment and transmission delay, therefore a 0.25 ms TTI provides the best performance. However, when evaluating the 99% percentile, it is observed that the achieved latency is considerably affected by the load. At low offered load, the optimal TTI size is 0.25 ms. However, as the load increases, the lowest latency is obtained with longer TTI size. Particularly, the 0.5 ms TTI provides equal or better performance for offered loads of 10 Mbps or higher.

TABLE II: Default simulation assumptions

Parameter	Value
SINR distribution	3GPP Macro network with 500 m ISD [13]; Full load conditions
System numerology	16 OFDM symbols per 1 ms; 17.143 kHz subcarrier spacing; 12 subcarriers per PRB [12]
System bandwidth	10 MHz; Effective transmission bandwidth of ∼9 MHz (44 PRBs)
TTI size	0.25 ; 0.5 ; 1 ; 2 ms
Scheduling technique	Fixed TTI size for all types of traffic; FCFS scheduling for LLC with priority over MBB
Traffic model	MBB: Single user with full buffer traffic LLC: Poisson arrival process with 1 kB payload
LLC offered load	0.4 - 12 Mbps

TABLE III: MBB throughput for different TTI sizes, SINR and LLC offered load.

SINR	TTI	4 Mbps off. load		8 Mbps off. load		12 Mbps off. load	
		Throughput	Gain[1]	Throughput	Gain[1]	Throughput	Gain[1]
[dB]	[ms]	[Mbps]	[%]	[Mbps]	[%]	[Mbps]	[%]
-3	0.25	3.29	0	2.13	0	0.98	0
	0.5	3.62	10	2.38	12	1.13	15
	1	3.77	15	2.48	17	1.22	23
	2	3.84	17	2.50	18	1.18	20
0	0.25	6.14	0	4.00	0	1.86	0
	0.5	6.45	5	4.28	7	2.13	14
	1	6.58	7	4.36	9	2.14	15
	2	6.63	8	4.34	9	2.13	15
+3	0.25	10.14	0	6.62	0	3.13	0
	0.5	10.44	3	6.92	5	3.38	8
	1	10.53	4	6.94	5	3.49	11
	2	10.55	4	6.99	6	3.38	8

[1]Gain relative to the 0.25 ms TTI configuration for the respective SINR and offered load parameters.

The main reason for this behaviour is the queuing delay. As the offered load increases, the queuing delay becomes the most dominant component on the total latency, therefore it is beneficial to increase the spectral efficiency (by using a longer TTI) in order to reduce the experienced delay in the queue. This phenomenon is illustrated in Fig. 5, where the queuing probability is plotted for different loads and TTI sizes. The queuing probability is defined as the probability that a user is not scheduled in the TTI immediately after arrival. It can be observed that, the shorter the TTI the higher the probability of experiencing queuing. Note that only a few cases experience a queuing probability higher than 50%, which reconfirms the steady behaviour of the observed median latency performance. Fig. 6 shows the distribution of the queuing delay for an offered load of 12 Mbps. It is observed that a 0.25 ms TTI configuration experiences the highest queuing delay (in both mean and tail of the distribution) as a consequence of the lower spectral efficiency. Configurations with 0.5 or 1 ms TTI provide lower queuing delay. At high load, the benefits of lower queuing delay exceed the drawbacks of longer transmission time and frame alignment delay, which results in overall better 99% percentile latency performance (as shown in Fig. 4).

The tradeoff between the TTI size and the queuing delay is not only evident when increasing the load, but also when

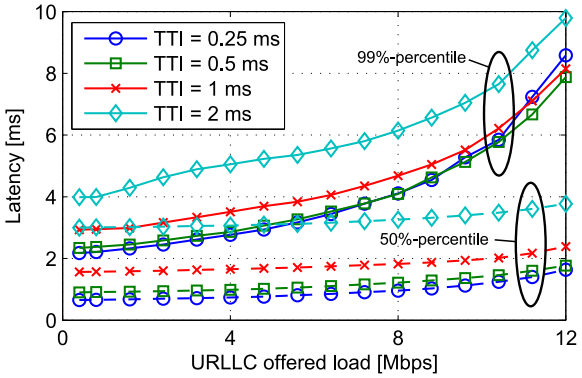

Fig. 4: LLC latency at the 50% and 99% percentile under different load conditions and TTI sizes.

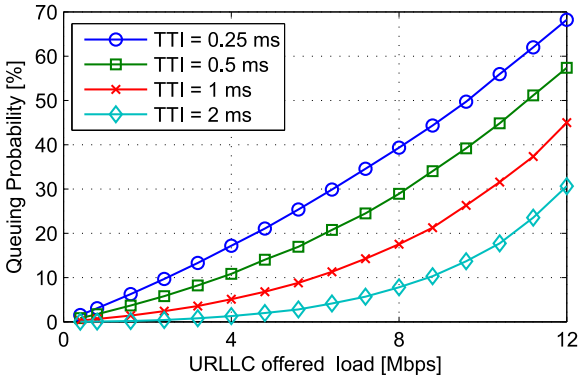

Fig. 5: LLC queuing probability under different load conditions and TTI sizes.

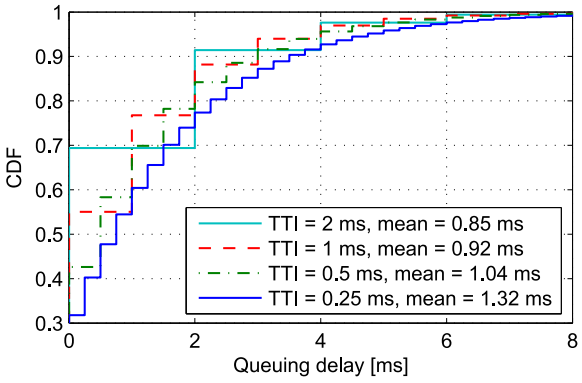

Fig. 6: Cumulative distribution function (CDF) of LLC queuing delay at 12 Mbps LLC offered load.

analysing the tail of the latency distribution. Fig. 7 shows the latency distribution for offered loads of 4 Mbps and 12 Mbps. For these two cases, we have run longer simulations such that it allows us to examine with good accuracy up to the 99.99% percentile. Even at relatively low load (4 Mbps, Fig. 7(a)), there is a gain from using a 0.5 ms TTI over a 0.25 ms TTI if the percentile of interest is above 99%. A similar trend is observed for the high load case (12 Mbps, Fig. 7(b)). However, in this case the point at which the 0.5 ms TTI becomes better than the 0.25 ms TTI appears much earlier in the distribution. It is also observed that the 1 ms TTI configuration is the best performing solution for percentiles above 99.9%.

VI. Discussion

The presented results show the benefits of using different TTI sizes to achieve low latency, depending on the offered load and the percentile of interest. Particularly, the tail of the latency distribution reveals the importance of using long TTI size (e.g. 0.5 or 1 ms) with higher spectral efficiency in order to reduce the experienced queuing delay. The observed trends are relevant for URLLC use cases, which require latencies of a few milliseconds guaranteed with reliability levels up to 99.999% [2]-[4]. However, the advantages of using different TTI sizes are broader. For example, the TTI duration can be adjusted in accordance to the user-specific radio channel conditions in order to compensate for the control overhead. This benefit has been shown in Table III, where the throughput gains of having large TTIs are more significant for users with low SINR. The TTI size can also be selected according to the individual user's service requirements. Besides URLLC and MBB, another relevant 5G use case is low cost massive machine-type of communication (mMTC) which might only support narrow bandwidth operation and therefore will benefit from long TTIs [2]. Given these manifold benefits, it is expected that a highly flexible scheduling of users, such as illustrated in Fig. 2, will be of key importance to efficiently support the different use cases and requirements envisioned for 5G.

VII. Conclusions

In this paper we have analysed the latency performance with different TTI configurations taking into account the multi-user dynamics of a cellular network. At low offered loads, it is observed how a 0.25 ms TTI is an attractive solution to achieve low latency. The main benefits come from the low frame alignment and transmission delay required to transmit the payloads. However, as the load increases, it has been shown how longer TTI sizes, e.g. 0.5 ms or 1 ms, provide improved performance as these configurations can better cope with the non-negligible queuing delay. The presented results allow to conclude that support for scheduling with different TTI sizes is important to achieve low latency and should be included in future 5G. Our future work will include a more detailed modelling of physical and medium access layer mechanisms including link adaptation, transmission errors and the respective retransmissions. Evaluations with simultaneous use of different TTI size depending on the use case is also of interest.

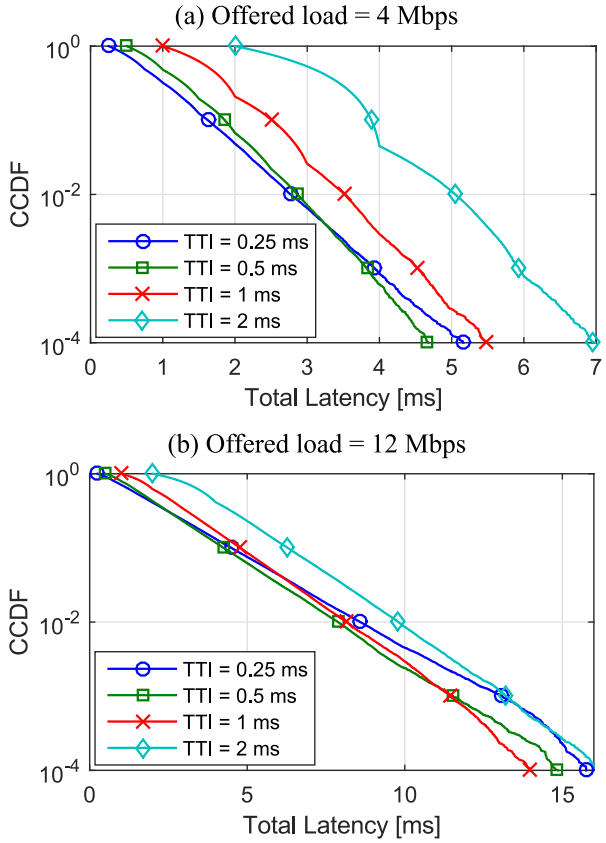

Fig. 7: Complementary cumulative distribution function (CCDF) of the LLC latency under different load conditions and TTI sizes.

References

[1] A. Osseiran et al., "Scenarios for the 5G mobile and wireless communications: the vision of the METIS project", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26-35, May 2014.

[2] 3GPP TR 38.913 v0.3.0, "Study on scenarios and requirements for next generation access technologies", March 2016.

[3] ITU-R M.2083-0, "IMT vision - framework and overall objectives of the future development of IMT for 2020 and beyond", Sept. 2015.

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

[5] S. Ahmadi, "LTE-Advanced: A practical systems approach to understanding 3GPP LTE releases 10 and 11 radio access technologies", Academic Press, 2013.

[6] F. Ishizaki and G. U. Hwang, "Queuing delay analysis for packet schedulers with/without multiuser diversity over a fading channel", *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 3220-3227, Sept. 2007.

[7] I. Rubin and Z.-H. Tsai, "Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems", *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 637-647, May 1989.

[8] G. Wunder and C. Zhou, "Queueing analysis for the OFDMA downlink: throughput regions, delay and exponential backlog bounds", *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 871-881, Feb. 2009.

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

[10] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks", *IEEE Globecom*, Dec. 2014.

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

[12] G. Berardinelli, K. I. Pedersen, F. Frederiksen and P. Mogensen, "On the design of a radio numerology for 5G wide area", *International Conference on Wireless and Mobile Communications*, Oct. 2015.

[13] 3GPP TR 36.814 v9.0.0, "Evolved universal terrestrial radio access (E-UTRA); Further advancements for E-UTRA physical layer aspects", March 2010.