Latency reduction for narrowband URLLC networks: a performance evaluation

Zubair Amjad1,2 · Kofi Atta Nsiah2,3 · Benoît Hilt2 · Jean-Philippe Lauffenburger2 · Axel Sikora1,3

Accepted: 16 January 2021 / Published online: 29 March 2021 © The Author(s) 2021

Abstract
Fifth-generation (5G) cellular mobile networks are expected to support mission-critical low latency applications in addition to mobile broadband services, where fourth-generation (4G) cellular networks are unable to support Ultra-Reliable Low Latency Communication (URLLC). However, it might be interesting to understand which latency requirements can be met with both 4G and 5G networks. In this paper, we discuss (1) the components contributing to the latency of cellular networks and (2) evaluate control-plane and user-plane latencies for current-generation narrowband cellular networks and point out the potential improvements to reduce the latency of these networks, (3) present, implement and evaluate latency reduction techniques for latency-critical applications. The two elements we detected, namely the short transmission time interval and the semi-persistent scheduling are very promising as they allow to shorten the delay to processing received information both into the control and data planes. We then analyze the potential of latency reduction techniques for URLLC applications. To this end, we develop these techniques into the long term evolution (LTE) module of ns-3 simulator and then evaluate the performance of the proposed techniques into two different application fields: industrial automation and intelligent transportation systems. Our detailed evaluation results from simulations indicate that LTE can satisfy the low-latency requirements for a large choice of use cases in each field.

Keywords Cellular networks · CIoT · URLLC · Low-latency · Industry automation · V2X

1 Introduction
In recent years, cellular networks have gained much attention and importance for Internet of Things (IoT) applications, as they promise low cost for installation and maintenance, deployment flexibility, and scalability for automation applications [1]. Next-generation cellular networks are being developed around three use cases, as presented in Fig. 1 [2]. In addition to the enhanced Mobile Broad-Band (eMBB), next generation cellular communication applications are divided into two categories [3]: massive Machine Type Communication (mMTC), and Ultra-Reliable Low-Latency Communication (URLLC). These applications are categorized based on the throughput, latency, reliability, and scalability of the network. While mMTC involves large number of low-cost devices with high requirements on battery life and scalability such as smart city and smart home, URLLC targets mission critical applications with stringent requirements on latency and reliability [4].
Industrial automation and Vehicle-to-Anything (V2X) communication are two major use cases of URLLC [5, 6]. Recently, in the context of Industry 4.0, the importance of cellular networks for industrial automation applications have been highlighted by the automation industry as well as the research community [7]. As the existing standardized wireless technologies do not fulfill the latency and reliability requirements [1], industrial automation applications mainly rely on wired fieldbus standards, like PROFINET, HART, and CAN [8] and cannot benefit from wireless communication technologies. V2X communication is one of the key technologies in Intelligent Transportation Systems (ITS), providing wireless connectivity between cars, infrastructure elements, and pedestrians [9]. For nearly a decade, Dedicated Short Range Communication (DSRC) based on IEEE 802.11p has been studied, and it appeared as a promising wireless technology for local V2X communications [10–12]. However, recent studies [13–15] show that the evolution of cellular networks has enabled Long Term Evolution (LTE) to become the preferred communication technology for V2X.

In the different releases of 3rd Generation Partnership Project (3GPP) which is a standards organization that develops protocols for mobile telephony, no new candidate technologies have been identified in 5G for the Machine Type Communication (MTC) application that could surpass the existing 4G narrowband LTE to meet the key requirements for better coverage, cost sensitivity, and battery longevity [16]. As a result, 3GPP has opted to improve the narrowband LTE system as a solution to delivering the MTC application in 5G. Consequently, the three application pillars of 5G (see Fig. 1), delivered by one 5G standard, actually require that the LTE systems also remain in place even with some further improvements to support mMTC and URLLC use cases. Therefore, improving the existing 4G LTE standards for supporting MTC is also essential.

The 4G LTE cellular networks have been continuously improved and standardized by the 3GPP. The major improvements till 3GPP Release 13 were focused on Mobile Broad-Band (MBB) applications. Therefore, the LTE user-plane latency remained unchanged from 3GPP Release 8 till Release 13 and does not meet the requirements of URLLC use cases [17]. To enable support for latency-critical application, in its Release 13 [18], 3GPP proposed latency reduction techniques for the next generation of low-latency cellular networks. 3GPP additionally standardized two narrowband User Equipment (UE) categories for MTC as Cat-M1 (LTE-M) and Narrowband-IoT (NB-IoT) in Release 13 [19] to support narrowband cellular communication providing better coverage, longer battery life and lower manufacturing cost for devices. The potential of latency reduction techniques is yet to be explored for narrowband cellular networks and, therefore, we
investigate these techniques for URLLC use cases (industry automation and V2X). Moreover, the narrowband UE categories are mainly considered for latency-tolerant application and their use for ultra-low latency application is in general neglected. We believe that there is a potential in low-cost, low-energy UE categories (such as LTE-M and NB-IoT) to fulfill the URLLC requirements for cellular networks.

In this paper, we explore the potential of latency reduction techniques for narrowband cellular networks. We first discuss the factors that contribute in the uplink latency of 4G LTE networks and then evaluate the latency by means of both theoretical calculations and simulations. The downlink latency is not considered in this work since the downlink latency in LTE networks is lower than 10 ms. We then explain latency reduction techniques – initially proposed by 3GPP [18] – and discuss their potential for decreasing latency in cellular networks. There is a lack of implementation and evaluation of these techniques in the known open-source simulators. To this end, we implement and evaluate latency reduction techniques for the narrowband UE category LTE-M in Medium Access Control (MAC) and Physical (PHY) layers of the LTE module [20] of the open-source Network Simulator (ns-3) [21]. In order to evaluate these techniques, the overall setup of the simulation and our development is shown in Fig. 2. The developed techniques are part of LTE module and the simulation scenarios are defined as part of the simulation setup which includes the definition of node topology, network parameters and configurations. The setup is then used to run the simulation of LTE module. The simulation model generates the trace files for each layer of LTE protocol stack containing different Key Performance Indicators (KPI). The results are then extracted from the trace files generated from the simulations. Our main contributions in this paper are as follows:

- Investigation of delay components in 4G LTE networks,
- In-depth analysis of minimum achievable uplink latency of cellular networks for narrowband UE by theoretical calculations and evaluation through realistic simulations,
- Conceptualization, implementation and evaluation of latency reduction techniques for narrowband cellular networks in the open source ns-3 simulator. The techniques are simulated for realistic URLLC use case scenarios.

The remainder of the paper is structured as follows: Sect. 2 presents the state of the art for latency reduction in narrowband cellular networks. In Sect. 3, we discuss the components of LTE uplink latency and compare the simulation results with theoretical calculations in Sect. 4. Afterwards, two latency reduction techniques, short transmission time interval and semi-persistent scheduling are discussed and their evaluation along with a discussion on the obtained results are presented in Sects. 5 and 6 respectively. Section 7 concludes the paper.

### 2 State of the art on latency reduction

Cellular network standards have been continuously evolved by 3GPP through different releases. As part of the evolution of fourth generation LTE networks, different latency reduction techniques have been described and evaluated in 3GPP study and work items [18, 22, 23]. We have mapped some of these techniques to the different parts of the cellular network architecture as shown in Fig. 3. For example, mobile edge computing, which enables offloading of Radio Access Network (RAN) and Evolved Packet Core (EPC) functionalities in powerful computing units in close proximity to the UEs, overlaps between the RAN and EPC. On the other hand, short Transmission Time Interval (sTTI) overlaps between the UE and evolved NodeB (eNB) and enables shorter transmission times between both. In the following, we outline the research works available in the literature that analyze the potential of latency reduction techniques for upcoming 5G cellular networks. All of these research works proposed and evaluated latency reduction techniques for wideband cellular networks, and therefore, there is a gap in the literature for the evaluation of latency reduction techniques for narrowband UE categories. It is important to evaluate latency reduction techniques for narrowband UE categories due to the limited bandwidth of the system, which can impact the improvements offered by these techniques.

Semi-Persistent Scheduling (SPS) [24] and Fast Uplink Access (FUA) [25] targeting latency reduction on the MAC
layer were the first steps for latency reduction in 3GPP Release 14 [22]. On the PHY layer, sTTI and reduced processing time belong to the second step towards latency reduction included in Release 15. The performance of sTTI and SPS as potential techniques to support latency-critical use cases is evaluated in our previous works [26, 27]. Transmission time interval lengths of two Orthogonal Frequency Division Multiplexing (OFDM) symbols (2-os) and 7-os together with 14-os legacy TTI are used in this work. Simulation was performed with the open-source ns-3 simulator and the benefits of the techniques are explored in the uplink for LTE-M category of UEs. The results show that short TTI and SPS greatly reduce the latency and can support mission-critical applications. However, the evaluation of these techniques for realistic industrial automation and V2X deployment scenarios is yet to be explored. In this paper, we extend our previous works [24, 27] by evaluating the sTTI and SPS techniques for two most common URLLC use cases (i.e. V2X and industry automation).

In [28], the authors conduct a link-level performance analysis of low latency operation in the downlink and uplink of LTE networks with different shortened TTI lengths of 1-os, 2-os, 7-os and the legacy 14-os. The evaluation of sTTI for narrowband category of UEs is not covered in their work. The authors in [29] conduct system-level performance analysis of the potential benefits of the sTTI and reduced processing time techniques for the reduction of the downlink and uplink latency as well as the Round Trip Time (RTT) latency in LTE networks. Short TTI of 2-os and 7-os are used in their simulation and are evaluated for eMBB use cases. Their results show that short TTI outperforms reduced processing time with respect to uplink, downlink and RTT latency. The focus of their work is the reduction of latency for eMBB use cases. Therefore, latency reduction for URLLC use cases is not covered. Performance evaluation of the combined short TTI and reduced processing time techniques are missing. Additionally, techniques for narrowband category of UEs are not evaluated.

The work in [30] proposes a combination of 7-os TTI together with the reduction of uplink access delay. The performance analysis of the system latency of the different TTI lengths and their proposed technique is evaluated for cell center and cell edge users by running different simulation scenarios. Their work focuses on improving latency for cell center and cell end users; however, the evaluation of their work concerns mainly the wideband and do not address the narrowband UE category. Two latency reduction techniques, reducing processing time and shortening the TTI were studied separately in [31]. In order to achieve a comparatively smaller latency, short TTIs of length 1-os, 2-os, 7-os, 14-os were used in their work. Comparing these TTI lengths, 1-os offers the best performance in terms of latency reduction; however, this introduces a significant overhead. Evaluation of the technique for narrowband UEs is not considered in this work.

Performance improvement of short TTI is also evaluated in the downlink of LTE networks in [32]. In their simulations, the TTI of 2-os, 4-os, 7-os and 14-os were used. Their results show that the shortest TTI configuration used in their work (i.e. 2-os) offers the best performance with regard to latency reduction at low traffic loads. Two areas are considered in this work, the evaluation of short TTI in the uplink of LTE and the evaluation of the technique for narrowband category of UEs. The authors in [33] evaluate short TTI and SPS for industrial automation applications to
identify the set of applications that could be supported with the latency reduction enhancements. They evaluated sTTI and SPS in system level in-house simulator over industry automation parameters. However, their work lacks the evaluation of latency reduction techniques for narrowband UE category.

Summarizing the contributions previously described, we classify the sTTI and SPS techniques and present in Table 1 the different delay components that have been manipulated. This shows clearly that the evaluations of latency reduction techniques were conducted for broadband UE categories. In addition, the evaluations were conducted using either in-house simulators or theoretical assumptions. Moreover, simulation evaluation of latency reduction techniques for realistic use case scenarios is also not present in the literature. To fill this gap, in this paper we perform a full analysis and optimization of sTTI and SPS for narrowband UE categories and evaluate our work with a widely used realistic and open source simulator for which we developed the latency reduction techniques.

3 Latency of 4G LTE: theoretical analysis

There are multiple components that contribute to the uplink latency in LTE network. These components impact directly the performance of the system and make it a challenge to meet the requirements of latency-critical use cases. In this section, we present an in-depth latency analysis for 4G LTE network and discuss the requirement for improvements towards 5G networks for URLLC applications. The presented analysis includes both the control-plane and user-plane latencies. For the control-plane latency, we investigate the Random Access (RA) procedure for narrowband UEs. The random access procedure enables the UE to acquire network resources and send/receive data. The user-plane latency analysis consists of theoretical calculations for each factor having a direct impact on the uplink latency. The user-plane latency is considered as the delay between the Packet Data Convergence Protocol (PDCP) layers of eNB and UE. In this paper, we do not investigate the downlink latency, since it is under 10 ms anyhow, as explained in [33]. The following sections bring the material for a good understanding of these components.

3.1 Control-plane latency: random access

To get access to the network resources, a ‘contention-based’ RA is performed by the UE on a dedicated physical channel called Physical Random Access Channel (PRACH). The PRACH consists of six Physical Resource Blocks (PRBs) and its location in the network resources is periodically transmitted by the eNB in a System Information Block 2 (SIB-2). A UE initiates the RA procedure whenever (a) it has new data to transmit but without uplink synchronization, (b) it recovers from a link failure, (c) it changes its state from Radio Resource Control idle (RRC_IDLE) to RRC_CONNECTED, or (d) it performs a handover.

Contention-based RA consists of four messages presented in Fig. 4 [34]. The first message is called preamble transmission, where a UE selects a random sequence from a pool of pre-defined preambles/sequences and transmits it over PRACH. A collision in preamble transmission occurs when two or more UEs send identical preamble simultaneously. In the normal case (i.e. no collision), eNB answers with a Random Access Response (RAR) to UE on Downlink Shared Channel (DLSCH). In the next phase, UE transmits a RRC message and initiates a contention resolution timer. Finally, if eNB receives the RRC connection request successfully, it signals UE about the completion of RA procedure with a contention resolution message. If the contention resolution timer expires, UE starts the RA procedure again from the preamble transmission phase. The duration of the RA procedure is a major contributor in LTE control-plane latency.

| Paper Refs. | BW (MHz) | Target use case | Evaluation platform | NB evaluation | Use case evaluation | Short TTI | SPS |
|-------------|---------|-----------------|---------------------|---------------|---------------------|-----------|-----|
| [26]        | 1.4     | URLLC           | Open-source         | ✔             |                    | ✔         | ✔   |
| [27]        | 1.4     | URLLC           | Open-source         | ✔             |                    | ✔         | ✔   |
| [28]        | 10      | –               | In-house            | ×             |                    | ×         | ×   |
| [29]        | 10      | eMBB            | In-house            | ×             |                    | ×         | ×   |
| [30]        | 20      | –               | In-house            | ×             |                    | ×         | ×   |
| [31]        | 20      | URLLC           | Theory              | ×             |                    | ×         | ×   |
| [32]        | 10      | URLLC           | In-house            | ×             |                    | ✔         | ×   |
| [33]        | 10      | URLLC           | In-house            | ×             |                    | ✔         | ✔   |

BW, Bandwidth; NB, Narrowband; Open-source simulator, ns-3
3.2 User-plane latency components

Uplink latency in LTE depends on multiple factors as presented in Fig. 5 and their corresponding numbers in Table 2. Most of the latency is related to message transmission time and processing time. The UE performs random access procedure before sending/receiving data from the network. The eNB allocates uplink resources whenever a UE requests the resources to send/receive data.

The frame structure of LTE is based on subframe with a duration of 1 ms and 14 symbols (see Fig. 7). The default TTI is 1 ms and the minimum processing time to encode/decode the received data either at eNB or UE is 3 ms [17]. These factors lead to an average uplink latency of 11.5 ms in LTE under ideal conditions (see Table 2). The user-plane latency components are briefly described in the following sections.

Table 2 Uplink latency components for 1 ms transmission time interval (14-os) [17, 35]

| Component                                | Delay (ms) |
|------------------------------------------|------------|
| Average delay to next SR opportunity     | 0.5        |
| SR transmission                          | 1          |
| eNB processing delay                     | 3          |
| Transmission of scheduling grant         | 1          |
| UE processing delay                      | 3          |
| Average delay to scheduled resources     | 0.5        |
| Transmission of uplink data              | 1          |
| eNB processing delay                     | 1.5        |
| Total uplink latency                     | 11.5       |
3.2.1 Grant acquisition

A UE must send Scheduling Request (SR) on Physical Uplink Control Channel (PUCCH) when it has data to send. In order to send a SR, the UE must wait for the PUCCH SR-valid resource and a Scheduling Grant (SG) in response to the SR. The UE can start transmitting data over Physical Uplink Shared Channel (PUSCH) after decoding the SG. The wait for the PUCCH resources and transmission/reception of SR/SG includes a latency of roughly 6 ms (see Table 2).

3.2.2 Transmission time interval

The transmission time interval refers to the transmission duration on the radio link. In LTE, a system frame has a length of 10 ms and consists of 10 sub-frames of 1 ms each. An illustration of transmission time interval is shown in Fig. 7. The transmission of a request, grant, and/or data is done in 1 ms sub-frames. Each transmission in LTE consists of 1 ms subframe, which is one of the major contributors in the uplink latency.

3.2.3 Processing and core

The processing time, either from UE or eNB, is the major contributor in LTE uplink latency. When a data unit is received at eNB or UE, it is processed in each layer starting from PHY till the application layer. The processing of each data unit also incurs extra delay, which is normally considered to be 3 times the TTI [17]. Both data and control messages need to be encoded/decoded in the network elements which adds to the end-to-end delay. Congestion in the Core Network (CN) due to packet queues can insert additional delay in the system performance. However, CN and Internet delays can vary widely.

4 Latency evaluation of 4G LTE

In this section, we analyze the control and user-plane latencies of narrowband LTE networks through simulation. With the aim to investigate narrowband UE category for MTC, we select LTE-M with the bandwidth of 1.4 MHz for the simulations. Due to lower cost, higher range and longer battery life, LTE-M is one of the preferred choices for MTC applications. Narrow bandwidth with a higher number of devices in the network pose a challenge in terms of latency requirements for the network. In the following section, we explain the simulations setup and evaluate the obtained results.

4.1 Simulation setup

The parameters for our simulations are shown in Table 3. In our previous work [26], we evaluated control-plane latency of LTE-M UE category with a realistic RA model. Further details about realistic ns-3 RA model can be found in [34]. Especially, the authors developed a RA model for ns-3 LTE module and evaluated it for wideband UE category. Moreover, we used their model to evaluate narrowband UE category for MTC in [26] by investigating multiple parameters of random access. In this paper, we further extend the latency analysis by evaluating the user-plane latency. The results presented in this paper are mean values calculated from ten independent simulation runs.

4.2 User-plane latency

The uplink user-plane latency is measured from the instance the data is passed to the LTE protocol stack of the UE until the moment the data is processed by the LTE protocol stack in the eNB. Figure 6 presents the comparison of LTE uplink latency for LTE-M (1.4 MHz bandwidth) and Cat-0 (20 MHz bandwidth) from simulations with an increasing number of devices (from 20 to 180) in the network. The uplink latency remains higher than 10 ms whatever the number of UEs in the network. For LTE-M, the latency increases slowly up to 60 UEs and then increases abruptly. This sudden increase happens when the available system capacity in terms of data rate is lower than the total data traffic offered from UEs. Furthermore, the increase in latency after resource saturation is linear because the scheduling in our simulations is performed in a round robin process. Therefore, the impact of increased

| Parameter                             | Value         |
|---------------------------------------|---------------|
| Simulator - version                   | NS - 3.26     |
| Propagation loss model                | Two-Ray Ground|
| eNB transmission power                | 43 dBm        |
| UE transmission power                 | 20 dBm        |
| Downlink EARFCN                       | 100           |
| Uplink EARFCN                         | 18100         |
| Uplink bandwidth                      | 1.4 MHz       |
| Number of resource blocks             | 6             |
| eNB noise figure                      | 5 dB          |
| UE noise figure                       | 9 dB          |
| Transmission time interval            | 1 ms          |
| Resource scheduling                   | Dynamic       |
| Simulation duration                   | 30 s          |

Table 3 Simulation parameters used in our ns-3 simulations for the evaluation of control-plane and user-plane uplink latency
number of UEs includes a linear latency increase. The latency for Cat-0 UEs also follows the same linear increase after resource saturation. The performance of different scheduling algorithms has been already investigated by Dawaliby et al. [36]. The maximum uplink data rate defined for LTE-M is 1 Mbps. The total data traffic offered to the network depends on the number of device in the network cell, and the application data rate which is used to send data to the network. In Fig. 6, each UE sends a data packet of 12 bytes every 100 ms, which results in a data rate of 9.6 kbps per UE. In case of LTE-M, the total bandwidth offered to the network by 100 UEs is 960 kbps, which is less than the peak data rate of LTE-M [36]. The data rate offered by 140 UEs is 1.34 Mbps, which exceeds the network capacity and overloads resources. LTE-M with 1.4 MHz bandwidth and 1 Mbps peak uplink data rate is clearly not suitable for high data rate (>1 Mbps) applications; however, it provides a promising potential for low data rate applications that require very low latency due to lower cost and higher range. In the following section, we explain some of the latency reduction techniques and evaluate these techniques through simulations of LTE-M UE category.

5 Analysis of latency reduction techniques

The control-plane and user-plane latency analyses in the previous section highlight the need for reduction in 4G cellular networks latency. To meet the requirements of latency-critical applications, 3GPP proposed latency reduction techniques in Release 13 [18]. In this section, we briefly outline two of those techniques and later evaluate them for URLLC use cases. The improvements from these techniques target both user-plane latency in both uplink and downlink. In this paper, we focus only on the evaluation for user-plane latency reduction in narrowband cellular networks. The control-plane latency comes mainly from the RA procedure (see Sect. 3.1), which is a bottleneck problem in uplink latency. However, control-plane latency can still be avoided in applications, where devices only perform the RA procedure once in the start and make periodic transmissions later. The UEs remaining in connected state are only affected by user-plane latency.

5.1 Short transmission time interval

The TTI duration in LTE adds to 1 ms with two slots (0.5 ms each) comprising 14 symbols (0.0714 ms each). With a reduction of the TTI length (i.e. less than 14-os), the overall data transmission and processing time can be significantly reduced [29, 37]. As presented in [38], sTTI of 2-os (i.e. 0.14 ms) and 7-os (i.e. 0.5 ms) shall be standardized in Release 15 to be supported by both downlink (DL) and uplink (UL) transmissions. The sTTI implies shorter time duration for transmission (see Fig. 7) and faster processing for data decoding. For instance, using a 2-os TTI can reduce the uplink latency from 11.5 ms to 2.36 ms [39].

The sTTI can lead to a backward compatibility issue, as all the control channels are designed for only 1 ms legacy TTI. For instance, DL control channel in legacy LTE occupies 1-3 symbols, which would need modifications to enable sTTI of two symbols. UEs supporting sTTI shall be able to coexist with legacy UEs within the same system bandwidth. This can be achieved by a flexible frame structure allowing both legacy and shortened TTI to be used by UEs for transmitting. Furthermore, sTTI can also lead to a decreased throughput, as with sTTI the amount of resources required for control channels also increases [38]. The control channels, such as Physical Downlink Control Channel (PDCCH) and Physical Broadcast Channel (PBCH) occupy two and three OFDM symbols in LTE resource grid respectively. Therefore, in order to achieve the same network performance as with 14-os, these channels are split into multiple transmission time intervals, reducing the total shared channel resources [32].

5.2 Semi-persistent scheduling

In LTE, the coordination of the radio resources and channel access is handled by the eNB itself. LTE systems usually use Dynamic Scheduling (DS) to maximize resource utilization. With DS, UEs send scheduling requests to get
access to the network resources for uplink data transmissions. In Voice over LTE (VoLTE) applications, persistent or semi-persistent scheduling is used to eliminate the extra scheduling overhead. However, for narrowband LTE UE categories, such as LTE-M and NB-IoT, the only scheduling type for data transmissions is dynamic that includes extra overhead for uplink transmissions. As depicted in Fig. 8, a UE needs to get access to uplink resources by first sending a scheduling request. The eNB then allocates resources for this UE through a scheduling grant. As presented in [17], with the legacy TTI of 1 ms, the uplink latency in LTE networks always remains above 12 ms (see Table 5 for details). There have been some recent studies [40–42] to improve scheduling in cellular networks, however, these studies do not target improving latency.

Semi-Persistent Scheduling (SPS) overcomes the extra control messages delay by eliminating the scheduling messages and associated processing duration. With SPS, the eNB schedules uplink transmissions for the UEs in connected state without the reception of scheduling request. As shown in Fig. 8, a UE does not have to wait for scheduling request/grant and can transmit the data as soon as it is ready to be sent out. With a default TTI of 1 ms,
SPS can reduce the uplink latency to 4.5 ms with the cost of reduced capacity due to the resource pre-allocation [24]. The system capacity is reduced with SPS because the regularly scheduled resources for UEs are not always utilized, particularly by UEs that do not have any data to send in every allocated resource (padding transmission in right side of Fig. 8). This method incurs two issues into the system, a) spectral inefficiency and b) increased energy consumption. Both issues are due to the undesired padding information sent to eNB when there is no data available to be sent.

6 Evaluation of latency reduction techniques

In this section, we present the results from simulations of sTTI and SPS techniques we implemented into the ns-3 LTE module. Indeed, the publicly available implementation of the LTE module does not support any of the latency reduction techniques. A sTTI feature for LTE module to include sTTI of 0.5 ms (7-os) and 0.14 ms (2-os) along with the legacy TTI of 1 ms (14-os) has been developed in this work. In our previous works [24, 27], latency reduction techniques with different numbers of UEs in a single network cell have been evaluated. Here, we extend our analysis for specific URLLC use cases taking account of realistic channel propagation models.

6.1 Simulation setup

The parameters used in our simulations for this section are shown in Table 4. For the evaluation of latency reduction techniques, simulations were performed for 3GPP LTE-M UE category with a bandwidth of 1.4 MHz (6 PRBs). In order to evaluate potential improvements through sTTI and SPS, we simulate three URLLC scenarios. In the first part, we consider a single stationary UE in the network and compare simulated latency with theoretical calculations.

In the second part, parameters for industry automation has been defined and an evaluation of the latency reduction techniques is done. We use Hybrid Building propagation loss model from ns-3 for realistic simulation and place the UEs inside a large hall in a building. The simulations are targeted towards evaluating process automation, where UEs send data periodically, and factory automation, where UEs send data based on an event. The trade-offs between device density and uplink latency are also discussed.

In the third part, we evaluate latency reduction techniques for V2X use case with mobile UEs over a small sized city map with another realistic propagation loss model for this specific application, i.e. the Three Log-Distance propagation model. In this paper, we only focus on network-assisted (communication between mobile UEs and the infrastructure) V2X communication. The discussion about Device-to-Device (D2D) communication is out of the scope of this paper.

6.2 Evaluation for a single UE

Theoretical calculations of uplink, downlink and end-to-end latency with sTTI and SPS are given in Table 5. A comparison between theoretical and simulated uplink latencies for a single UE is shown in Fig. 9 for 2-os, 7-os and the legacy 14-os TTI with dynamic and semi-persistent scheduling. These results validate the features developed in ns-3 LTE module. The small difference in simulated and theoretical calculation is due to the waiting time for the

| Parameter                          | Value                      |
|------------------------------------|----------------------------|
| Simulator - version                | NS - 3.26 with sTTI and SPS extension |
| Propagation loss model             | Three log-distance / Hybrid building |
| eNB transmission power             | 43 dBm                     |
| UE transmission power              | 20 dBm                     |
| Downlink EARFCN                    | 100                        |
| Uplink EARFCN                      | 18100                      |
| Uplink bandwidth                   | 1.4 MHz                    |
| Number of resource blocks          | 6                          |
| eNB noise figure                   | 5 dB                       |
| UE noise figure                    | 9 dB                       |
| Packet size                        | 50 B                       |
| Packet transmission interval       | 100 ms                     |
| Transmission time interval         | {1, 0.5, 0.14} ms          |
| Resource scheduling                | Dynamic / Semi-persistent  |
| Simulation duration                | 30 s                       |
resources to send either scheduling request or data. As presented in Table 2, the average waiting time to resources considered for theoretical calculations is 0.5 ms; however, it could be between 0 to 1 ms in simulations.

It is worth noticing that with 14-os TTI and dynamic scheduling, the minimum uplink latency is theoretically limited to 11.5 ms. Multiple transmissions required to complete a scheduling request/grant procedure (see Table 2) are the main cause for the minimum latency limit for 14-os TTI. In this case, each uplink and downlink transmission adds a delay of 1 ms. However, for shorter TTI, both the transmission and the processing times decrease, which results in a reduced overall latency. Moreover, SPS removes the necessity of scheduling request/grant messages, which further reduces the latency. The short TTI of 2-os together with SPS can reduce the uplink latency for a single device for more than 85% from the baseline legacy TTI with dynamic scheduling.

### 6.3 Evaluation for industry automation

We extend the evaluation analysis of latency reduction techniques for industry automation applications. The extended evaluation includes realistic channel propagation model and simulation scenario. Figure 10 shows the uplink latency for a varying number of UEs in a factory hall depicting the scenario of industry automation. All the UEs are considered to be stationary in the simulations and periodically send data to the eNB. All UEs communicate with a single eNB installed in the factory hall. The periodic data transmissions represent mainly the process automation, where sensors on the machines send data periodically. It is important to note from the results that short TTI can satisfy the 10 ms latency requirement for narrowband UE category with a limited number of UEs (i.e. <30) in a cell. The major factors impacting the latency of narrowband LTE are number of UEs, packet size, and packet sending interval. The uplink latency of 50 ms is still achievable in this use case for 60 UEs in a single cell.

---

**Table 5** Theoretical latency calculations for different lengths of short TTI together with dynamic or semi-persistent scheduling. The end-to-end latency includes only the access network delay

| Feature | TTI (ms) | Processing time (ms) | Scheduling | UL latency (ms) | DL latency (ms) | E2E latency (ms) |
|---------|----------|----------------------|------------|----------------|----------------|------------------|
| 14-os   | 1        | 3                    | Dynamic    | 11.5           | 4.5            | 16               |
|         |          |                      | SPS        | 5.5            | 4.5            | 10               |
| 7-os    | 0.5      | 2                    | Dynamic    | 5.5            | 2.5            | 8                |
|         |          |                      | SPS        | 3.5            | 2.5            | 6                |
| 2-os    | 0.14     | 1                    | Dynamic    | 2.7            | 0.93           | 3.63             |
|         |          |                      | SPS        | 1.44           | 0.93           | 2.37             |

**Fig. 9** Uplink latency comparison of theoretical and simulated values for three different TTI and two different scheduling schemes with a single UE sending data to a unique cell in eNB. TTI of 2-os with SPS reduces the latency for approximately 85% as compared to the baseline legacy 14-os TTI with dynamic scheduling (DS)
We further extend our analysis of sTTI and SPS to discuss the trade-offs for satisfying URLLC latency requirements. This time, we evaluate factory automation (event-triggered data uploads) and process automation (periodic data uploads) with different number of UEs in the network cell. The rationale behind this evaluation is to discuss the limit on the number of UEs while fulfilling the URLLC latency requirements. Figure 11 represents the uplink latency for different number of UEs with periodic data uploads. The periodic data transmissions represent mainly the process automation use case, where sensors on the machines send data periodically. The latency of legacy 14-os TTI with dynamic scheduling is always more than 10 ms, while shorter TTI (i.e. 2-os) with dynamic scheduling manages to keep the latency under 5 ms. The latency of 7-os DS remains under 10 ms. The latency increase with SPS for ten or more UEs is due to the fact that the network resources are allocated for transmissions to all the UEs even if there is no data to send with the UEs. Therefore, SPS is less effective in case of higher number of UEs (i.e. [10]) in a cell.

The evaluation of the factory automation use case, where UEs send data only once based on an event is presented in Fig. 12. The latency remains below 10ms with short TTI (2-os and 7-os) for 14 UEs in a cell, however, dynamic scheduling manages to keep the latency below 5 ms for 18 UEs. In both applications i.e. factory automation and process automation, a latency below 10 ms is guaranteed under the condition, that the network resources are not saturated and the UEs are in connected state i.e. they do not need to perform RA procedure again. Therefore, narrowband UE categories can fulfill the URLLC latency requirements while trading-off between maximum number of devices in the cell, packet size and packet sending interval.

6.4 Evaluation for V2X

The simulation evaluation of sTTI for V2X use case, where all the UEs are mobile (i.e. cars) is presented in Fig. 13. The simulations were performed on the map of Offenburg city (Germany) with a realistic channel propagation model and four eNBs installed at different locations in the city. The mobility of UEs and realistic channel propagation model affect the uplink latency. It can be noticed from the results that shorter TTI and SPS cannot fulfill 10 ms requirement for a comparatively larger number of UEs (i.e. >30). However, in the simulated scenario, sTTI and SPS prove to be very effective in keeping the uplink latency under 10 ms for 25 or less mobile UEs. The uplink latency comparison of different TTI for the industry automation use case, where multiple static UEs send data periodically to a single eNB

![Fig. 10 Uplink latency comparison of different TTI for the industry automation use case, where multiple static UEs send data periodically to a single eNB](image)
Fig. 11 Uplink latency comparison for UEs sending data periodically to the eNB with different TTI and scheduling for industry automation use case. The main purpose of this evaluation is to analyze the number of UEs in a cell for which the network can fulfill the latency requirements. Short TTI is clearly a better choice over the legacy 14-os TTI. The short TTI can further reduce the latency to 2 ms with the help of SPS but only for a very small number of devices.

Fig. 12 Uplink latency comparison of event-triggered data transmissions with different TTIs and scheduling techniques for industry automation use case, where the UEs send the data only once during the simulation representing an alarm or a malfunction. This evaluation is targeted towards analyzing the limits on number of UEs in a cell for which the network can meet the latency requirements. 2-os TTI outperforms legacy TTI with both type of scheduling. However, dynamic scheduling performs better in this case due to much sophisticated resource utilization.

Latency for 30 UEs in Fig. 13 is higher than 10 ms. It is also important to mention here that reduction in TTI leads to an increase in control overhead, which obviously affects resource utilization. As compared to the industry automation use case simulation in Fig. 10, the number of active eNBs is four times larger in V2X simulations due to the larger coverage area. This is also the reason behind very low difference in uplink latency for both use cases.

The short TTI and SPS can significantly reduce latency at the cost of higher control overhead and resource utilization efficiency. Moreover, enabling support for multiple-sized TTI within the same network cell is very important and needs further evaluation. Obviously, the UEs that can only support legacy TTI cannot utilize sTTI, and therefore, providing backward compatibility for TTI is a necessary step towards enabling ultra-low latency 5G cellular networks.
7 Conclusion

One of the newer requirements for cellular networks is to enable the support for mission-critical IoT applications. The increasing demand from such applications with regard to latency poses a challenge for cellular networks. In this paper, we presented an overview of the state of the art and URLLC use cases and their requirements, analyzed and evaluated the control and user-plane latency of 4G cellular networks, analyzed the short TTI and SPS latency reduction techniques and evaluated their potential in supporting latency-critical applications. The evaluation covered industrial automation applications with periodic as well as event-triggered traffic patterns, and V2X applications with periodic data uploads. We used the TTI lengths of 2-os, 7-os and legacy 14-os together with dynamic and semi-persistent scheduling. We then implemented TTI length variants and SPS in the open-source ns-3 simulator and evaluated these techniques for narrowband LTE-M category of UEs. Results show that, for a single UE, short TTI of 2-os with SPS reduces the latency for more than 85% compared to the legacy TTI 14-os with dynamic scheduling. With an increased number of UEs, where some UEs send data periodically and others send data sporadically, the results show that 2-os short TTI with either SPS or DS can significantly reduce the latency. Thus, these combinations have the potential to support URLLC applications with stringent latency requirements.

As an extension to the work presented in this paper, we plan to implement and evaluate the short TTI and SPS techniques for Cat-NB1 category of UEs and also to investigate the challenges of enabling simultaneously different-sized transmission time interval in a network cell. The ns-3 code will be made public once our ongoing development activities are complete. Moreover, as a next step to this work, the control overhead incurred by reducing the transmission time interval will also be investigated.

Acknowledgements This work was partly funded by the Région Alsace, France and the Institute of Reliable Embedded Systems and Communications Electronics (ivESK) at Offenburg University of Applied Sciences. The work done in this paper is part of the Taktius project partially funded by Federal Ministry of Education and Research, Germany, under 16K1S0812. The authors are grateful for this support.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Funding Open Access funding enabled and organized by Projekt DEAL.
References

1. Froitzsch, A., Wetzker, U., Bauer, M., Rentschler, M., Beyer, M., Elspass, S., & Klessig, H. (2014). Requirements and current solutions of wireless communication in industrial automation. In 2014 IEEE international conference on communications workshops (ICC) (pp. 67–72). IEEE.

2. ITU-R. (2017). Minimum requirements related to technical performance for IMT-2020 radio interfaces. Technical Report: ITU - International Telecommunication Union.

3. 3GPP TR 38.913 v15.0.0. (2018). Study on scenarios and requirements for next generation access technologies.

4. Dahlman, E., Mildh, G., Parkvall, S., Peisa, J., Sachs, J., Selén, Y., et al. (2014). 5G wireless access: Requirements and realization. IEEE Communications Magazine, 52(12), 42–47.

5. Li, Z., Uusitalo, M. A., Shariatmadari, H., & Singh, B. (2018). 5G: The wireless access for the Internet of Things. IEEE Communications Magazine, 56(6), 44–49.

6. ITU-R. (2016). Technical Report: ITU - International Telecommunication Union.

7. Ashraf, S. A., Aktas, I., Eriksson, E., Helmersson, K. W., & Ansari, J. (2016). A highly scalable IEEE802.11p traffic scheduling. In 2016 15th international symposium on wireless communication systems (ISWCS) (pp. 1–6). IEEE.

8. Thomesse, J. P. (2005). Fieldbus technology and industrial automation. In 2005 IEEE conference on emerging technologies and factory automation (Vol. 1, pp. 651–653). IEEE.

9. Abboud, K., Hassan, A. O., & Weihsu, Z. (2016). Interworking of DSRC and cellular network technologies for V2X communications: A survey. IEEE Transactions on Vehicular Technology, 65(12), 9457–9470.

10. Luoto, P., Bennis, M., Pirinen, P., Samarakoon, S., Horneman, K., & Latva-aho, M. (2017) Vehicle clustering for improving enhanced LTE-V2X network performance. In 2017 European conference on networks and communications (EuCNC), (pp. 1–5). IEEE.

11. Sikora, A., & Manuel S. (2013). A highly scalable IEEE802.11p communication and localization subsystem for autonomous urban driving. In: 2013 international conference on connected vehicles and Expo (ICCVE). IEEE.

12. Ledy, J., Poussard, A. M., Vauzelle, R., Hilt, B., & Boeglin, H. (2012). AODV enhancements in a realistic VANET context. In: 2012 international conference on wireless communications in unusual and confined areas (ICWCUCA), (pp. 1–5). IEEE.

13. Chen, S., Hu, J., Shi, Y., Peng, Y., Fang, J., Zhao, R., et al. (2017). Vehicle-to-everything (v2x) services supported by LTE-based systems and 5G. IEEE Communications Standards Magazine, 1(2), 70–76.

14. Chen, S., Hu, J., Shi, Y., & Zhao, L. (2016). LTE-V: A TD-LTE-Based V2X solution for future vehicular network. IEEE Internet of Things Journal, 3(6), 997–1005.

15. Okyama, S. (April). Intelligent transport systems towards automated vehicles. In ITU News Magazine (pp. 29–32).

16. Bi, Qi. (2019). Ten trends in the cellular industry and an outlook on 6G. IEEE Communications Magazine, 57(12), 31–36.

17. Takeda, K., Wang, L. H., & Nagata, S. (2017). Latency reduction toward 5G. IEEE Wireless Communications, 24(3), 2–4.

18. 3GPP TR 36.881 v0.6.0. (2016). Evolved Universal Terrestrial Radio Access (E-UTRA); Study on Latency Reduction Techniques for LTE.

19. 3GPP TR 45.820. (2015). Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT).

20. LTE-EPC Network Simulator. (2019). http://networks.cttc.es/mobile-networks/software-tools/lena. Accessed 19.

21. 3GPP TR 18.190667. (2016). Work item on L2 latency reduction techniques for LTE.

22. 3GPP TR 18.191299. (2016). Work item on shortened TTI and processing time for LTE.

23. Amjad, Z., Sikora, A., Hilt, B., & Lauffenburger, J. P. (2018). Latency Reduction for Narrowband LTE with Semi-Persistent Scheduling. In 2018 IEEE 4th international symposium on wireless systems within the international conferences on intelligent data acquisition and advanced computing systems (IDAACS-SWS) (pp. 196–198). IEEE.

24. 3GPP TR 18.191590. (2016). Work item on L2 latency reduction techniques for LTE.

25. Amjad, Z., Sikora, A., Lauffenburger, J. P., & Hilt, B. (2018). Low latency V2X applications and network requirements: Performance evaluation. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 220–225). IEEE.

26. Amjad, Z., Sikora, A., Lauffenburger, J. P., & Hilt, B. (2018). Latency reduction in narrowband 4G LTE networks. In 2018 15th International Symposium on Wireless Communication Systems (ISWCS) (pp. 1–5). IEEE.

27. Hosseini, K., Patel, S., Damnjanovic, A., Chen, W., & Montojo, J. (2016). Link-level analysis of low latency operation in LTE networks. In 2016 IEEE Global Communications Conference (GLOBECOM) (pp. 1–6). IEEE.

28. Arenas, J. C. S., Dudda, T., & Falconetti, L. (2017). Ultra-low latency in next generation LTE radio access. In SCC 2017; 11th International ITG Conference on Systems, Communications and Coding (pp. 1–6). VDE.

29. Xiaotong, S., Nan, H., & Naizheng, Z. (2016). Study on system latency reduction based on Shorten TTI. In 2016 IEEE 13th International Conference on Signal Processing (ICSP) (pp. 1293–1297). IEEE.

30. Zhang, X. (2017). Latency reduction with short processing time and short TTI length. In 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) (pp. 545–549). IEEE.

31. Zhang, Z., Gao, Y., Liu, Y., & Li, Z. (2018). Performance evaluation of shorted transmission time interval in LTE networks. In 2018 IEEE Wireless Communications and Networking Conference (WCNC) (pp. 1–5). IEEE.

32. Aktas, I., Jafari, M. H., Ansari, J., Dudda, T., Ashraf, S. A., & Arenas, J. C. (2017). LTE evolution-Latency reduction and reliability enhancements for wireless industrial automation. In 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC) (pp. 1–7). IEEE.

33. Polese, M., Centenaro, M., Zanella, A., & Zorzi, M. (2016). M2M massive access in LTE: RACH performance evaluation in a smart
city scenario. In 2016 IEEE International Conference on Communications (ICC) (pp. 1–6). IEEE.
35. 3GPP TR 36.912 v14.0.0. (2015). Feasibility study for Further Advancements for E-UTRA (LTE-Advanced).
36. Dawaliby, S., Bradai, A., & Pousset, Y. (2016). In depth performance evaluation of LTE-M for M2M communications. In 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (pp. 1–8). IEEE.
37. Pocovi, G., Pedersen, K. I., Soret, B., Lauridsen, M., & Mogensen, P. (2016). On the impact of multi-user traffic dynamics on low latency communications. In 2016 International Symposium on Wireless Communication Systems (ISWCS) (pp. 204–208). IEEE.
38. Li, J., Sahlin, H., & Wikstrom, G. (2017). Uplink phy design with shortened tti for latency reduction. In 2017 IEEE Wireless Communications and Networking Conference (WCNC) (pp. 1–5). IEEE.
39. Schulz, P., Matthe, M., Klessig, H., Simsek, M., Fettweis, G., Ansari, J., et al. (2017). Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture. IEEE Communications Magazine, 55(2), 70–78.
40. Adesh, N. D., & Renuka, A. (2019). Adaptive downlink packet scheduling in LTE networks based on queue monitoring. Wireless Networks, 25(6), 3149–3166.
41. Ragaleux, A., Baey, S., & Gueguen, C. (2016). Adaptive and generic scheduling scheme for LTE/LTE: A mobile networks. Wireless Networks, 22(8), 2753–2771.
42. Alavikia, Z., & Ghasemi, A. (2018). Overload control in the network domain of LTE/LTE: A based machine type communications. Wireless Networks, 24(1), 1–16.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Zubair Amjad received B.S. Telecommunication Engineering from National University of Computer and Emerging Sciences, Pakistan, and M.Eng. Computer Engineering from Jeju National University, South Korea, in 2013 and 2016 respectively. He was awarded a Ph.D. in Computer Science at University of Haute-Alsace, France with a thesis on Ultra-Reliable Low-Latency Communication in 5G cellular networks in 2020. His main areas of interest are URLLC, 5G, industry automation, predictive maintenance, and low latency wireless communication.

Kofi Atta Nsiah received B.Sc. Computer Science from Kwame Nkrumah University of Science and Technology, Ghana, and M.Sc. Communication Engineering from Offenburg University of Applied Sciences, Germany, in 2008 and 2016 respectively. His main areas of interest are 5G/6G wireless networks, Low latency wireless communication, millimeter communication, machine learning, Industrial Internet of Things.

Benoît Hilt received a Ph.D. in electrical engineering from the Institut National Polytechnique de Lorraine (France) in 1991. He joined the Network and Telecommunication Department of the University of Haute Alsace as associate professor in 1994. He is member of the Telecom and Network team of IRIMAS Lab. Since 2017 he is habilitated to lead research (HdR). His current research interests include vehicular networking and time sensitive networking.

Jean-Philippe Lauffenburger obtained a Ph.D. in electrical and control engineering at the Université de Haute Alsace (UHA), Mulhouse, France, in 2002. From 2004 to 2015, he was assistant professor at UHA in the ENSISA engineering college. Since 2016, he has been Full Professor in ENSISA where he is dean of the Control Engineering and Embedded Systems department and vice-dean of the international relations. His major research interests are data fusion applied to perception for intelligent transportation systems. He is the author or co-author of 20 international patents and publications in international journals. He is co-author of about 60 conference papers and supervised 10 Ph.D. students. Prof. Lauffenburger is a Member of the IFAC “Transportation Systems” TC7.4 and a Reviewer for the Information Fusion Journal, IEEE Transactions on Intelligent Transportation Systems or IEEE Transactions on Vehicular Technology. He is Associate Editor of the IEEE Transactions of Intelligent Vehicles journal and is a regular Associate Editor of IEEE and IFAC conferences.
Axel Sikora holds a diploma (Dipl.-Ing./M.Sc.) of Electrical Engineering and a diploma of Business Administration (Dipl. Wirt-Ing. / MBA), both from Aachen Technical University, Germany. He has done a Ph.D. (Dr.-Ing.) in Electrical Engineering at the Fraunhofer Institute of Microelectronics Circuits and Systems, Duisburg, with a thesis on SOI-technologies. After various positions in the telecommunications and semiconductor industry, he became a professor at the Baden-Wuerttemberg Cooperative State University Loerrach in 1999. In 2011, he joined Offenburg University of Applied Sciences, where he leads the Institute of Reliable Embedded Systems and Communication Electronics (ievSK). Since Jan 2016, he is also deputy member of the board to Hahn-Schickard Association of Applied Research, where he leads the division “Software Solutions”. Since October 2019, he is also affiliated to Technical Faculty of Freiburg University, allowing him to serve as primary Ph.D. supervisor.

His major interest is in the field of efficient, energy-aware, autonomous, and value-added algorithms and protocols for wired and wireless embedded communication.