LATENCY REDUCTION TOWARD 5G

KAZUKI TAKEDA, LI HUI WANG, AND SATOSHI NAGATA
NTT DOCOMO AND DOCOMO BEIJING COMMUNICATIONS LABORATORIES CO., LTD.

INTRODUCTION

A fifth generation (5G) mobile network system is required to realize latency much shorter than that of the current fourth generation (4G) mobile network systems, in order to be capable of supporting services with very low latency requirements including driverless cars, enhanced mobile cloud services, real-time traffic control optimization, emergency and disaster response, smart grid, e-health, efficient industrial communications, as well as tactile internet, augmented reality, factory automation, and so on.

Toward 5G, the Third Generation Partnership Project (3GPP) has identified two evolution paths and is working on them:
1. Enhancing Long Term Evolution (LTE)
2. Establishing the New Radio (NR) access technology

Currently, working on lower latency is ongoing for both LTE and NR in 3GPP Working Groups (WGs). Enablers for low latency are essential to meet the IMT-2020 requirements for a use case called ultra-reliable and low-latency communications (URLLC), which is identified as one of the cornerstones in 5G [1].

In this short article, we investigate 3GPP standardization works aimed at latency reduction by LTE and NR, and introduce certain technical challenges.

LATENCY IN LEGACY LTE

In LTE, a subframe is defined as 14 orthogonal frequency-division multiplexing (OFDM) symbols spanning a 1 ms duration [2]. The subframe is an actual and typical transmission time interval (TTI) of data. The latency is in general a function of the TTI. For example, LTE U-plane one-way latency is presented in [3] as shown in Table 1.

For FDD-LTE, t_{FA} = 0.5 ms and t_{HARQ_RTT} = 8 ms. Assuming the target block error rate (BLER) of data channel being P_{BLER} = 0 and 10 percent, the total one-way delay can be derived as 4.0 ms and 4.8 ms, respectively.

For uplink, typically, UE initiates the data scheduling by sending a scheduling request (SR) to the network. This incurs additional delay as following in Table 2.

For frequency-division duplex (FDD)-LTE, t_{HARQ_RTT} = 8 ms and the minimum value of t_{SR} = 0.5 ms by setting SR on all the subframes. Assuming the target BLER of a data channel being P_{BLER} = 0 and 10 percent, the uplink latency including scheduling delay can be derived as 10.5 ms and 11.3 ms, respectively.

Until Release 14, U-plane latency had been kept unchanged from Release 8. The requirements for 5G in URLLC services is 1 ms and 0.5 ms in IMT-2020 and 3GPP [1, 4], respectively, and none of them can be met by the current LTE.

LATENCY REDUCTION FOR LTE

For LTE, latency reduction techniques were studied in Rel. 14 [5]. Based on the study outcome, a layer 2 solution was specified in Rel. 14 [6], and a layer 1 solution is to be specified in Rel. 15 [7].

The layer 2 solution is to allow the user equipment (UE) to skip uplink transmission if the UE has no data. More specifically, a network can configure/schedule uplink resources for a UE without taking into account data buffer of the UE, and then the UE can decide whether to transmit or skip uplink data depending on whether data is available in the UE buffer. In the legacy LTE, the UE shall send data in response to an allocated UL dynamic or configured grant even if no data is available in the UE buffer. Allowing uplink transmission skipping decreases uplink interference and improves UE energy efficiency, and makes semi-persistent resource allocation more realistic.

Step	Description	Value
1	BS processing delay	1 ms
2	TTI alignment	t_{FA}
3	Transmission of DL data	1 ms
4	UE processing delay	1.5 ms
5	HARQ retransmission	$t_{HARQ\_RTT} \times P_{BLER}$
	Total one-way delay	3.5ms + t_{FA} + $t_{HARQ\_RTT} \times P_{BLER}$

TABLE 1. U-plane one-way latency.

Step	Description	Value
1	Ave. delay to next SR opportunity	t_{SR}
2	SR transmission	1 ms
3	BS decodes SR and prepares UL grant	3 ms
4	Transmission of UL grant	1 ms
5	UE decodes UL grant and prepares data	3 ms
6	Transmission of UL data	1 ms
7	BS processing delay	1.5 ms
8	HARQ retransmission	$t_{HARQ\_RTT} \times P_{BLER}$
	Total one-way delay	10.5ms + t_{SR} + $t_{HARQ\_RTT} \times P_{BLER}$

TABLE 2. Uplink latency including scheduling delay.

The layer 1 solution includes two sub-solutions:
1. Shortened processing time for 1 ms TTI
2. Shortened TTI with shortened processing time

For 1, minimum hybrid automatic repeat request acknowledgment (HARQ-ACK) feedback delay and uplink data scheduling delay are shortened from 3.5 ms + t_{FA} = 4 ms to 3 ms (in case of FDD-LTE) while keeping TTI duration and all the existing channel structures unchanged. This requires the UE and network to simply shorten processing timelines for data transmission/reception and HARQ, resulting in reducing U-plane latency by 25 percent. For 2, TTI lengths of 2 symbols (2-os) and 7 symbols (7-os) are to be specified for FDD-LTE (for time-division duplex [TDD]-LTE, shortened TTI length of 7-os only is supported), which correspond to reductions of 86 and 50 percent TTI length, respectively. In general, shorter TTI length enables shorter processing time. Therefore, 2 is expected to realize further latency reduction compared to 1.

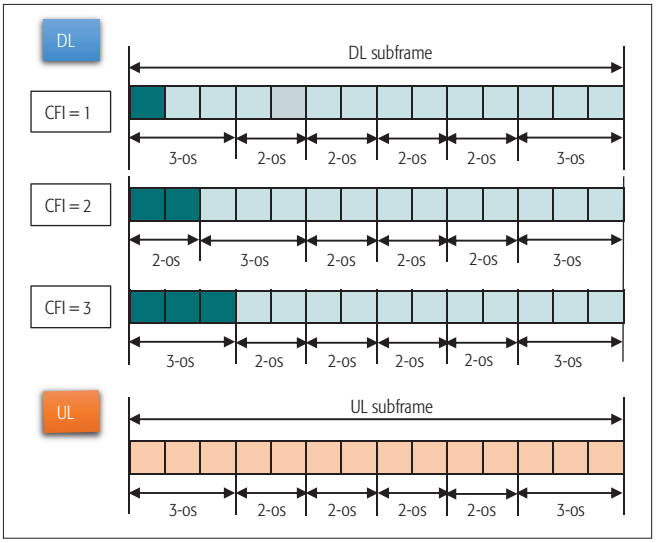

FIGURE 1. DL and UL short-TTI patterns for 2-symbol (2-os) sTTI.

One of the biggest technical challenges of LTE latency reduction, especially for the layer 1 solution, is to ensure backward compatibility. UEs supporting latency reduction shall be able to coexist with legacy UEs in the same serving cell. This requirement would restrict potential new designs of the layer 1 solution. One example is the layout of 2-os short-TTI. In the legacy LTE, DL control channel spans 1–3 OFDM symbols, which is informed to the UEs by the value of control format indicator (CFI) in a dynamic manner. Since the DL control channels for legacy UEs are interleaved and distributed over the whole control regions spanning the symbols, any data cannot be mapped in the control region. The sTTI layout was determined taking that into account. Besides, between the slots within a subframe, intra-subframe frequency hopping may be performed for legacy UEs; even in this situation, efficient resource allocation between the legacy TTI and sTTI is desirable. Furthermore, various reference signals (RSs, e.g., CRS, CSI-RS) specified so far are distributed within a subframe. Taking all these aspects into account, the layouts of 2-os short-TTI are determined [8] as following (Fig. 1):

- For downlink: {3, 2, 2, 2, 2, 3} for CFI=1 and for CFI = 3, {2, 3, 2, 2, 2, 3} for CFI = 2
- For uplink: {3, 2, 2, 2, 2, 3}

The above layouts ease the co-existence with legacy LTE TTI and various RSs.

As for backward compatibility, it is also important to make sure that the UE configured with latency reduction operation should not lose its coverage compared to legacy LTE. Theoretically, as the TTI length is shorter, the coverage for a given transmit power and a given payload is reduced. A typical example is the coverage of the uplink control channel (PUCCH) delivering a limited number of uplink control information (UCI) bits. Compared to legacy PUCCH, the PUCCH having 7-symbol without intra-TTI frequency hopping requires additional 5.9 dB to meet the performance requirement [9]. In order to ensure the coverage while achieving latency reduction benefit, support of dynamic fallback from short-TTI to legacy 1 ms TTI is necessary. However, changing TTI length from time to time dynamically complicates processing timelines and HARQ procedures, and creates another design challenge.

Assuming that the processing timeline is linearly shortened together with the TTI, 2-symbol short-TTI offers U-plane one-way latency of 0.8 ms in the case of 0 percent BLER, which can meet at least 1 ms delay budget defined as a latency requirement for IMT-2020.

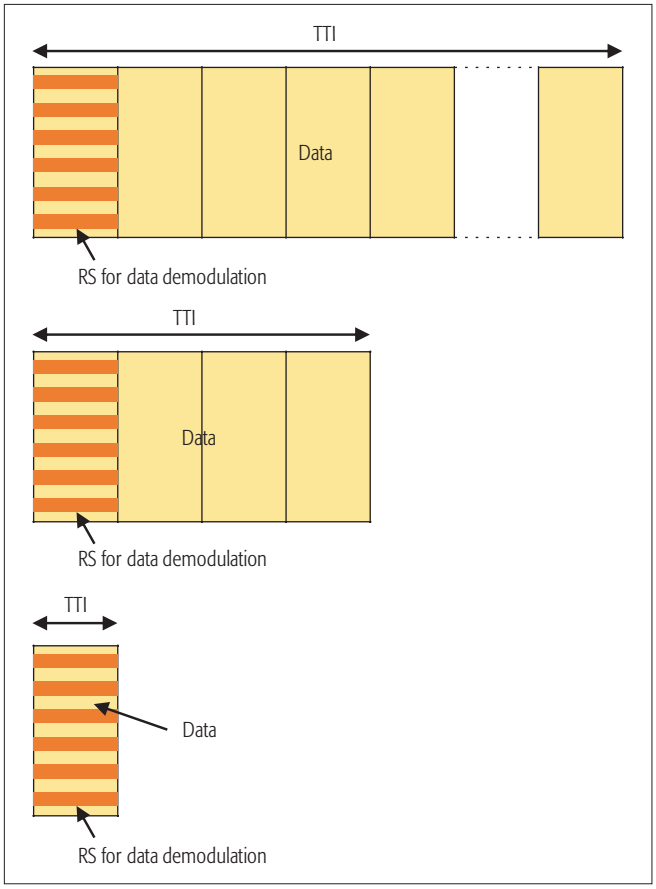

FIGURE 2. Example of data channel structure enabling flexible TTI durations.

NEW RADIO ACCESS TECHNOLOGY

For TTI shortening in NR, there are two approaches. The first approach is to allow data TTI having smaller number of OFDM symbols (e.g., 1 or 2), similar to LTE short-TTI. The difference from LTE short-TTI is that backward compatibility is not required for NR [10]. Therefore, it is possible to design a whole NR system, including control/data channels, reference signals, and related UE behaviors, from the beginning, such that various TTI lengths can flexibly be applicable depending on the service type of each UE. UEs with various TTI lengths shall be able to coexist in the same carrier efficiently. In order to realize such flexible TTI durations in a unified manner, reference signals for data demodulation should be confined within a limited number of OFDM symbols (e.g., 1 symbol). As long as the reference signals for demodulation is available, the number of OFDM symbols in which a data spans can be shorter (Fig. 2). Then, different TTIs having different numbers of OFDM symbols can be multiplexed on the same carrier in a flexible manner as illustrated in Fig. 3. This design principle ensures forward compatibility; that is, future new services requiring a certain data rate, latency, reliability, and so on can be realized by the framework.

Together with the flexible TTI duration, for NR, a new concept is now under consideration: sub-blocking of a TTI. More specifically, the transport block for a given TTI is divided into sub-blocks, and each sub-block is encoded and modulated

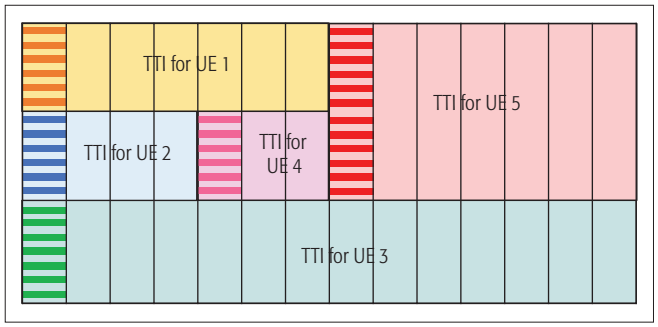

FIGURE 3. Multiplexing data channels having different TTI durations on one carrier.

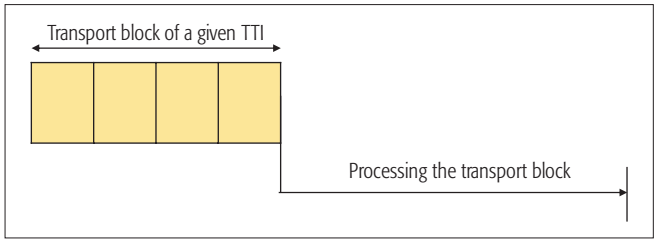

FIGURE 4. Processing timeline with non-pipeline processing.

independently. For LTE, such a mechanism is already there, called code-block segmentation, to reduce the burden on the decoder. Unlike legacy code-block segmentation, sub-blocks can physically be mapped on different OFDM symbols as much as possible. By this, the receiver can perform demodulating and decoding sub-blocks in a pipeline processing manner. Since one sub-block occupies a limited number of OFDM symbols, and different sub-blocks are mutually independent, the structure is interpreted as equivalent to concatenated multiple short TTIs, each having a sub-block. The resulting processing timeline can be decoupled by TTI duration, and can be much shorter than the TTI duration. Figures 4 and 5 show the comparison between process per transport block and process per sub-block. The extreme example is the self-contained structure in Fig. 6, in which case, within a given TTI duration, downlink data and its HARQ-ACK feedback channel are multiplexed in a time-division multiplexing (TDM) manner. This requires an OFDM symbol-level processing timeline. However, if the sub-block-based structure is enabled with 1 OFDM symbol granularity, such very quick processing and feedback may not be unrealistic.

Another approach is to use higher subcarrier spacing to shorten the OFDM symbol duration of a TTI. For example, 4 times higher subcarrier-spacing makes OFDM symbol duration in time to be 1/4. This can be an alternative to using a smaller number of OFDM symbols in a TTI. Multiplexing data transmissions modulated by different subcarrier- spacing may require some special handling. In the case of FDM, transmissions using different subcarrier spacing creates interference with each other. Therefore, either setting guard band or limiting modulation order/multiple-input multiple-output (MIMO) layers would be necessary. TDM does not create any additional interference and hence is easy to realize.

CONCLUSIONS

This column overviews the latency of a legacy LTE system, and introduces recent works on latency reduction in LTE and NR. For LTE, L2 and L1 solutions are identified and specified with keeping backward compatibility. For NR, the overall system is designed such that various TTI durations can be supported by a unified framework in order to ensure forward compatibility. These solutions will meet the requirements specified in IMT-2020 and 3GPP, and create URLLC services in the real world.

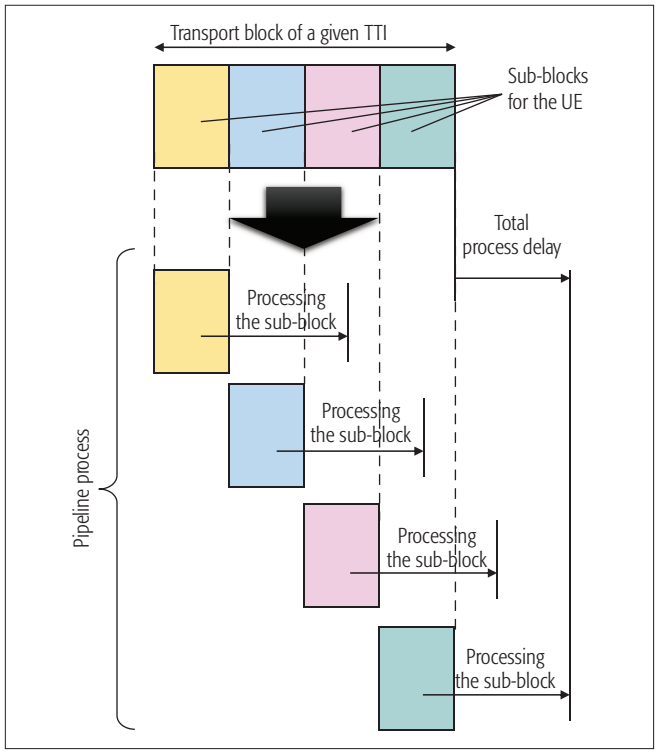

FIGURE 5. Processing timeline with pipeline processing.

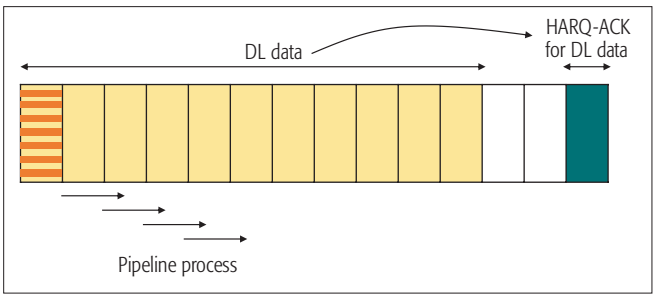

FIGURE 6. Self-contained TTI.

REFERENCES

[1] ITU-R Rec. M.2083, "IMT Vision — Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.
[2] 3GPP TS 36.211.
[3] 3GPP TR 36.912.
[4] 3GPP TR 38.913.
[5] 3GPP TR 36.881
[6] 3GPP RP-161747, "RAN2 Agreed CRs on Introduction of L2 Latency Reduction Techniques," Sept. 2016.
[7] 3GPP RP-170113, "Revised Work Item on Shortened TTI and Processing Time for LTE," Mar. 2017.
[8] 3GPP R1-1613691, "Way Forward on 2-Symbols DL sTTI Layout Structure," Intel, Ericsson, CATT, Nokia, Alcatel-Lucent Shanghai Bell, Huawei, HiSilicon, NTT DoCoMo, Samsung, Qualcomm, and Vodafone, Nov. 2016
[9] 3GPP R1-165212, "sPUCCH for Shortened TTI," NTT DoCoMo, May 2016.
[10] 3GPP RP-170847, "New WID on New Radio Access Technology," Mar. 2017.