International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792


# Energy Efficiency and Delay in 5G Ultra Reliable Low Latency Communications System Architectures

K. S. Deepa[1], S. P. Audline Beena[2], D. Rajini Girinath[3]
[1]*PG Student, Department of CSE, Sri Muthukumaran Institute of Technology, Chennai, India*
[2]*Assistant Professor, Department of CSE, Sri Muthukumaran Institute of Technology, Chennai, India*
[3]*Professor and HoD, Department of CSE, Sri Muthukumaran Institute of Technology, Chennai, India*

***Abstract*: Emerging 5G Ultra-Reliable Low-Latency Communications (URLLC) wireless systems are characterized by minimal over-the-air latency and stringent decoding error requirements. The low latency requirements can cause conflicts with 5G Energy Efficiency (EE) design targets. Therefore, it provides a perspective on various trade-offs between energy efficiency and user plane delay for upcoming URLLC systems. For network infrastructure EE, we propose solutions that optimize base station on-off switching and distributed access network architectures. For URLLC devices, we advocate solutions that optimize EE of Discontinuous Reception (DRX), mobility measurements, and the handover process, respectively, without compromising on delay.**

***Keywords*: Ultra-reliable low-latency communications (URLLC), 5G.**

## 1. Introduction

Ultra-Reliable Low-Latency Communications (URLLC) is one of the cornerstones of the upcoming fifth generation (5G) New Radio (NR) cellular system framework, together with Enhanced Mobile Broadband (eMBB) and Massive Machine Type Communications (mMTC). The key requirements of URLLC as per the Third Generation Partnership Project (3GPP) are to minimize the over-the-air latency of user plane data (at most 0.5 ms on average), while simultaneously ensuring very high packet reception reliability (error rates of at most 10–5). These constraints are expected to be critical for cutting-edge network applications such as augmented/virtual reality, autonomous ground vehicles, industrial Internet of Things (IoT) applications such as factory automation, pilotless aircraft, and remote surgery, to name a few. A rule-of-thumb comparison of the typical data transmission latencies and error rates for various connectivity protocols is third generation (3G) systems such as wideband code-division multiple access (WCDMA) are still in use today but are optimized for voice and low data rates, and latencies are especially increased when multiple users are multiplexed in the code domain. Fourth generation (4G) Long Term Evolution (LTE) offers improvements in over the- air latency, but cannot achieve URLLC reliability. Narrowband IoT (NB-IoT) and enhanced machine type communications (eMTC) protocols are designed to optimize energy efficiency of low-bandwidth devices, but cannot simultaneously provide low latency since they make extensive use of time-domain repetitions for coverage enhancement. It is seen that NR URLLC lies in a hitherto unexplored region between existing 3G/4G wireless standards and wire line protocols such as Ethernet (IEEE 802.3). The 3GPP URLLC standardization and academic studies have therefore been focused on the NR physical layer design needed to achieve the latency and reliability criteria. The interplay of URLLC latency and energy efficiency (EE) has received less attention. For example, initial studies have been performed on delay-aware downlink scheduling algorithms. While EE aspects of 5G eMBB systems have been studied previously, the latency criterion of URLLC invites further analysis. From a system perspective, network infrastructure EE and device or user equipment (UE) EE are equally important. About 80 percent of a mobile network's energy is consumed by base station sites, and carbon emissions from network infrastructure account for over 2 percent of the global total. On the other hand, a typical approach for increasing EE is to reduce the transmission or reception durations of network nodes in order to conserve power, which tends to increase packet delays. Therefore, improving the EE of a URLLC radio access network (RAN) without compromising on latency is an important consideration for the upcoming 5G ecosystem. The endeavor of this article is to explore the emerging URLLC system architecture and some of the associated trade-offs between delay and EE that have not yet been addressed in the standardization process. An overview of NR URLLC and the significance of EE is provided in the following section. A discussion of three aspects of network infrastructure EE is then presented along with corresponding solutions. Case studies in device EE are addressed following that. The proposed solutions may be employed individually or in combination, depending on the specific needs of the network deployment.

### *A. URLLC overview*

URLLC requirements cannot be met with existing 4G access

technologies such as Release 14 LTE, since the minimum transmission time interval (TTI) is 1 ms1 and the typical data packet errorrate target is 10–1. Furthermore, uplink (UL) LTE transmissions generally follow a three-step sequence of:

- Scheduling request on UL
- UL grant from eNB
- UL transmission after several TTIs

This series of events takes at least 8ms. Therefore, a new design and scheduling approach is necessary for NR URLLC. The NR air interface is based on cyclic prefix- orthogonal frequency-division multiplexing (CP-OFDM) as in LTE. However, multiple OFDM subcarrier spacings are supported ([15, 30, 60, 120, 240] kHz) as opposed to the 15 kHz used for LTE data and control channels. An NR URLLC transmission can be created by allocating a large bandwidth for the data and using an OFDM numerology with short symbol durations. Furthermore, a TTI in NR can be as short as two OFDM symbols; a two-symbol transmission with 120 kHz subcarrier spacing would span (1/(120×103)) = 16.67ms in the time domain (excluding CP). An NR slot with normal CP comprises 14 OFDM symbols and can be used for either downlink (DL) or UL transmissions, thereby enhancing transmission flexibility compared to the fixed duplexing modes of LTE. A key feature in 5G NR is the utilization of large-scale antenna arrays, or so-called massive multiple-input multiple-output (MIMO) for advanced beam forming. This raises the question of whether larger antenna arrays require higher 5G Node B (gNB) power consumption due to DL reference signal transmissions. The continuous, omnidirectional transmission of wideband cell-specific reference signals (CRSs) every DL subframe in LTE is wasteful if there are no or few UEs attached to the cell. 5G NR tackles this by eliminating CRSs and using and using channel state information reference signals (CSI-RSs) instead for CSI measurements and demodulation reference signals for data decoding. While an LTE CRS is present every four OFDM symbols in each DL slot in the time domain, an NR CSI-RS is configured on between 1–4 OFDM symbols per slot every {5, 10, 20, 40, 640} slots. Thus, NR reference signals can be much sparser in the time domain, which aids EE. To truly reduce latency, it is imperative that a URLLC data packet be transmitted as soon as it is received at the gNB or base station on the DL, or generated by the UE on the UL. However, this implies that time-frequency resources are always available whenever URLLC data needs to be transmitted. This complicates DL and UL scheduling since resources may have already been allocated or be in use by regular eMBB traffic. The NR design solutions for this problem are based on preemption on the DL/UL and autonomous transmissions on the UL, respectively. The gNB preemptively inserts URLLC data and control traffic into a part of the DL resources that are currently in use for an eMBB transmission. In other words, some of the lower-priority eMBB data is overwritten by the URLLC transmission. eMBB UEs need to be informed of the puncturing so as to reduce the degradation of their packet decoding. A similar principle is applicable to the UL, where UEs with URLLC transmission can overwrite UL resources in use by eMBB UEs. On the UL, autonomous transmissions are another latency-reducing option, where URLLC UEs transmit on pre-defined UL resources without the need for an explicit grant from the gNB. This mechanism is a natural extension of the semi-persistent scheduling scheme in LTE, the difference being that in NR the UE does not transmit if its UL data buffer is empty. Note that many of the details of the NR URLLC air interface and procedures remain under discussion at this time. Finally, several higher-layer techniques have also been introduced for NR URLLC. One such example is UL packet duplication at the Packet Data Convergence Protocol (PDCP) layer, which implies that a UE with dual connectivity to an LTE and an NR base station can utilize resources on both links for the same UL data. This serves to increase reliability via frequency diversity. All such higher-layer measures will benefit from lower latency at the physical layer, which is the core focus of this work.

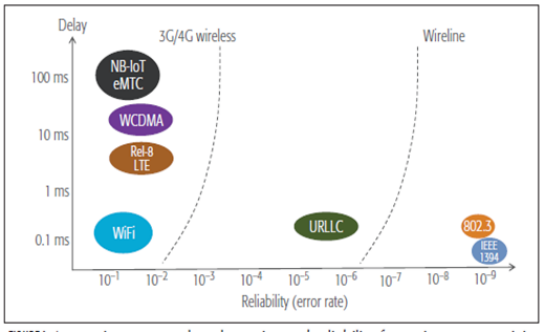


Fig. 1. Approximate user plane latencies and reliability for various connectivity protocols. Narrow band IoT and enhanced machine type communications are energy-efficient but have long repetition delays for coverage extension

## 2. Proposed system

### *A. On-off switching*

LTE was originally designed to have always-on DL transmissions from the eNB; specifically, certain wideband reference signals are transmitted every TTI. This leads to poor EE when there are no active UEs or no DL traffic to serve. The concept of evolved Node B (eNB) on-off switching was introduced in Release-12 as a remedy, where eNBs could suspend all transmissions for tens of milliseconds, without the need for handover of the served UEs to another eNB. The EE-delay trade-off is apparent when extending this concept to gNB on-off switching for URLLC: going into off mode can conserve energy, but leads to delays in delivering and receiving URLLC traffic. A potential solution is to utilize coordinated on-off switching across a set of adjacent gNBs. An example scenario is depicted in for the case of three coordinated gNBs. The gNBs share a sleep schedule among themselves, wherein gNBs with lower offered traffic and fewer connected UEs select longer OFF durations, in units of system frame numbers (SFNs), where

one frame spans 10 ms. The table in shows an example of such a coordinated sleep schedule, where gNB A is directed to go into off mode during SFNs, and so on.

*B. Advantages*

Reliability -Reliability is ensured by using very low-rate error correction coding together with multi-antenna beam forming. Energy Efficiency-We have seen so far that URLLC has stringent delay and reliability requirements. Energy efficiency has not been assigned explicitly as a performance metric for URLLC. Delay-Reception delay or latency in 4G and 5G systems can be divided into two major parts: user plane (UP) latency and control plane (C-Plane) latency.

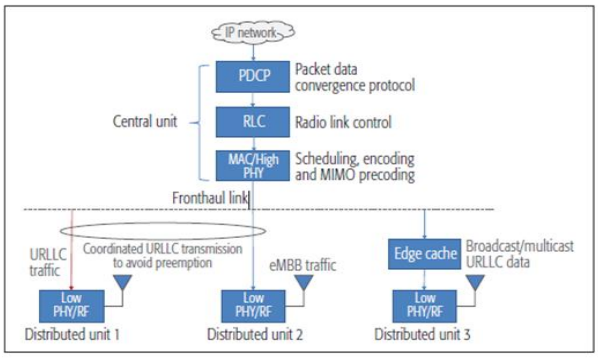


Fig. 2. EE strategies for distributed architectures

Frequency scanning for cell selection and measurements is another major cause of UE energy consumption. In LTE, the main mechanism to reduce UE power consumption in connected mode is to periodically send the UE to sleep. This is known as discontinuous reception (DRX), where the UE wakes up at pre-defined time instances (known as "on duration") to check for control channel transmissions directed to it. Power saving mode (PSM) is another LTE EE feature where the UE indicates to the network how often it needs to be active in order to transmit and receive data, entering a low-power state without DL monitoring in between. The network should not page the UE when it is in PSM, and moreover, should hold any DL data that arrives for the UE

## 3. System architecture

5G systems are being designed to be amenable to centralized or cloud RAN (CRAN) architectures with a functional split between a central units (CU) and multiple distributed units (DUs). Unlike traditional RANs, the baseband units (BBUs) for baseband processing are centralized in the CU as a BBU pool, leaving the front-end DUs with rudimentary filtering and signal processing. Each DU is configured only with the essential radio frequency components and some basic transmission/reception functionalities. The DUs are connected to the BBUs through high-bandwidth and low-latency front haul links. The global control of BBU processing at the CU leads to capacity and coordination efficiencies, particularly in terms of inter-cell interference mitigation. Separating the BBUs from the DUs can clearly lead to an increase in latency. The energy cost of preemption is also more pronounced, since additional energy is expended on transporting the punctured and potentially UN-decode able eMBB data to the DU over the fronthaul. Due to decoding failures, this data must then be retransmitted, which further degrades infrastructure and device EE. Consider two potential solutions for the CRAN case. The first builds on the gNB coordination principle used for on-off switching, and is appropriate for overlapping coverage scenarios such as in an industrial IoT setting. The CU routes URLLC traffic to whichever DU is currently not already serving eMBB data. The CU coordinates DU 1 and DU 2 in order to minimize preemption; URLLC data is served via DU 1 while eMBB traffic is served via DU 2. However, the front haul latency remains present in the system. Another solution is to deploy data caches in the system, preferably close to the network edge. A cache is a network entity configured to store and serve data; this reduces latency compared to fetching data all the way from the core network. An edge cache is deployed together with DU 3. A more comprehensive review of 5G caching strategies is presented. For the specific case of URLLC, caching is appropriate for broadcast and multicast data that must be served to multiple UEs. Note that gNB coordination and caching are complementary solutions that can be deployed together to further optimize the EE-delay trade-off.

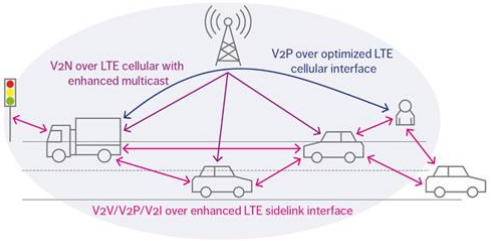


Fig. 3. System architecture

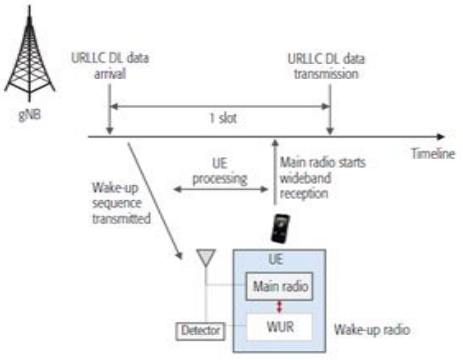


Fig. 4. DRX UE with a wake-up radio to minimize energy expended on wideband signal reception.

## 4. Results

The source eNB releases the UE resources (approximately 10ms).

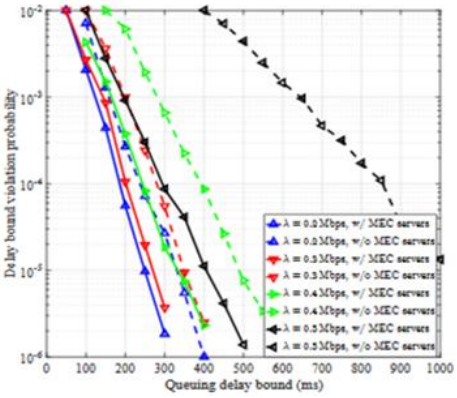


Fig. 5. Delay bound violation probability versus queuing delay bound

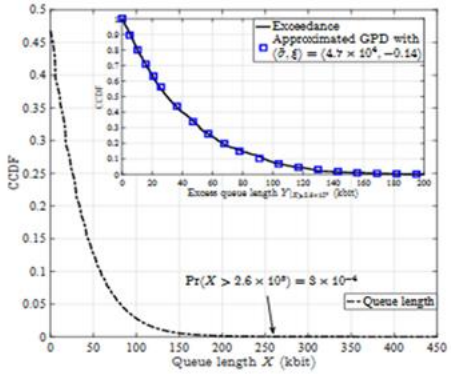


Fig. 6. Tail distributions of a given UE's task queue length, queue length exceedance over threshold, and the approximated GPD of exceedances

It is desirable to optimize the HO procedure in the case of URLLC for several reasons. During the measurement reporting and HO preparation phases, a URLLC UE will suffer from steadily degrading signal-to-noise ratio from its serving cell. This will impact the target reliability and diminish the quality of service. Second, LTE and NR both feature hard HOs wherein data transfers are interrupted until the completion phase. Minimizing the data interruption time is therefore vital for low-latency use cases. One solution to reduce HO latency is a "make before-break" HO mechanism where the UE attaches to the target gNB while still connected to the source. The drawback of this approach is that dual connectivity is required to both the source and target gNBs, which requires the presence of multiple RF chains at the UE for DL reception, together with advanced time-division multiplex switching capabilities on the UL. Another solution to address the above concerns is illustrated in Fig. 6b. The main change is for the URLLC UE to directly send a HO request to the target gNB based on its measurements. The role of the source gNB is bypassed, and if the target gNB accepts the request, data transfers from the new serving cell can begin more quickly. If the target gNB rejects the UE's request, the system falls back to the existing gNB-assisted HO.

## 5. Conclusion

This has touched upon the implications of various aspects of 5G URLLC systems with regard to energy efficiency and latency. The proposed solutions, which focus on the user plane and over-the-air delay, are a first attempt to address the associated trade-offs in the incipient NR system framework. Once the standards have matured, it would be worthwhile to also study backhaul, core network, and transport delays, and how these could be reduced via caching and network coding procedures. In particular, the detailed interplay of these delay parameters when combined with the distributed system architecture invites further scrutiny. Other important topics are the EE/delay aspects of the initial access procedure itself, joint EE optimization across network and UEs, connection reestablishment in the case of radio link failure, and cases where the NR traffic must coexist with LTE signals on the same carrier. It must also be ensured that the C-plane latency is not a bottleneck for URLLC performance. In conclusion, it is evident that 5G URLLC systems offer a rich variety of open research issues in terms of the trade-off between EE and delay.

## References

[1] 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies," June 2017.
[2] O. N. C. Yilmaz et al., "Analysis of Ultra-Reliable and Low-Latency 5G Communication for a Factory Automation Use Case," Proc. IEEE ICC Wksps, 2015.
[3] C. Sun, C. She, and C. Yang, "Energy-Efficient Resource Allocation for Ultra-Reliable and Low-Latency Communications,"Proc. IEEE GLOBECOM, 2017.
[4] H. Shariatmadari et al., "Optimized Transmission and Resource Allocation Strategies for Ultra-Reliable Communications,"Proc. IEEE PIMRC, 2016.
[5] H. Ji et al., "Introduction to Ultra Reliable and Low Latency Communications in 5G," 2017; https://arxiv.org/ abs/1704.05565.
[6] Nokia WP, "Building Zero-Emission Radio Access Networks,"2016.
[7] E. Dahlman, S. Parkvall, and J. Skold, 4G, LTE-Advanced Proand The Road to 5G, 3rd ed., 2016.
[8] 3GPP TS 38.211 V1.2.0, "NR; Physical Channels and Modulation (Release 15)," Nov. 2017.
[9] 3GPP TS 38.214 V1.1.2, "NR; Physical Layer Procedures for Data (Release 15)," Nov. 2017.
[10] M. Sybis et al., "Channel Coding for Ultra-Reliable Low-Latency Communication in 5G Systems," Proc. IEEE VTC-Fall2016, Montreal, Quebec, Canada, 2016, pp. 1–5.
[11] I. Parvez et al., "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," 2017; arXiv: 1708.02562v1.
[12] A. Mukherjee, "Queue-Aware Dynamic On/Off Switching of Small Cells in Dense Heterogeneous Networks," Proc. IEEE GLOBECOM Wksps, Dec. 2013.
[13] 3GPP TR 38.801, "Study on New Radio Access Technology: Radio Access Architecture and Interfaces," 2017.
[14] D. Zeng et al., "Take Renewable Energy into CRAN toward Green Wireless Access Networks," IEEE Network, no. 4, July2017, pp. 62–68.
[15] IEEE 802.11-16/1045r9, "A PAR Proposal for Wake-Up Radio," 2016.