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

Amitav Mukherjee

Abstract

Emerging 5G 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 EE design targets. Therefore, this work 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.

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) [1]. 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 [2–5].

A rule-of-thumb comparison of the typical data transmission latencies and error rates for various connectivity protocols is shown in Fig. 1. 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 wireline protocols such as Ethernet (IEEE 802.3). Meeting such stringent new requirements for a wireless access technology is one of the challenges of the ongoing NR design process that is expected to be complete by June 2018.

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 [3]. 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 [6]. 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. The article concludes with avenues for further research in the final section.

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 ms[1] and the typical data packet error

[1] Release 15 LTE will feature shortened TTIs of 0.2 ms duration, but without significant enhancements in reliability since the LTE channel coding framework will be reused. This aspect will be further enhanced in the LTE High Reliability and Low Latency work item in progress.

Amitav Mukherjee is with Verizon Communications.

Digital Object Identifier:
10.1109/MNET.2018.1700260

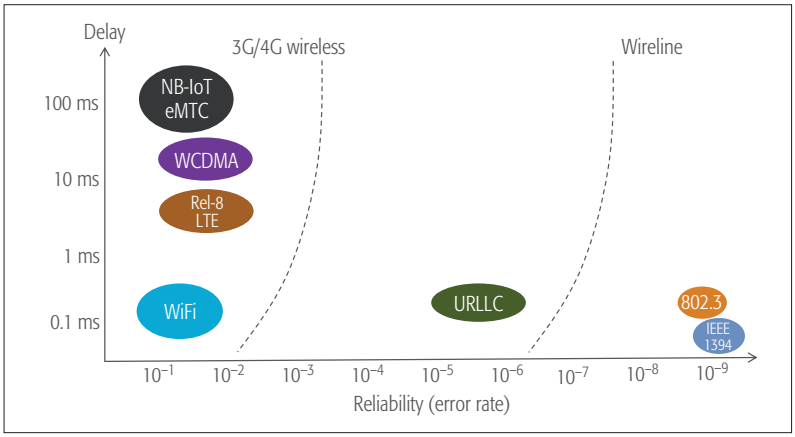

FIGURE 1. Approximate user plane latencies and reliability for various connectivity protocols. Narrowband IoT and enhanced machine type communications are energy-efficient but have long repetition delays for coverage extension. Note that WiFi has a higher variability due to operation in unlicensed spectrum.

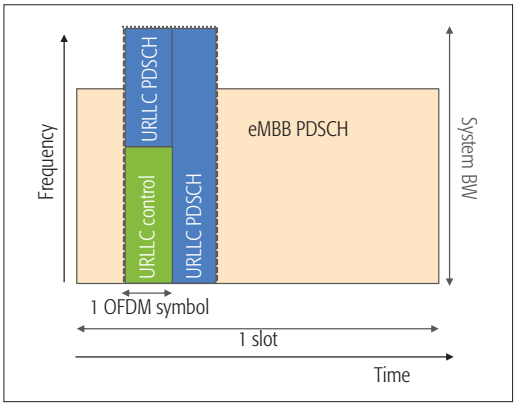

FIGURE 2. Preemption of eMBB physical downlink shared channel (PDSCH) with URLLC PDSCH. The URLLC data spans a larger frequency allocation than the eMBB transmission.

rate target is 10^{-1} [7]. 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 8 ms. 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 $(1/(120 \times 10^3)) = 16.67$ μs 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 beamforming. 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 [8]. 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 concept of preemption is illustrated in Fig. 2. Here, 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 [7]. 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 [9]. This mechanism is a natural extension of the semi-persistent scheduling scheme in LTE [7], 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.

Reliability

Reliability is ensured by using very low-rate error correction coding together with multi-antenna beamforming. In order to adhere to a short TTI, a typical URLLC data transmission occupies a large

fraction of the carrier bandwidth, which provides sufficient time-frequency resources for low-rate coding. Channel coding in NR makes use of new low-density parity check codes and polar codes for data and control channels, unlike the turbo codes used in LTE [10]. NR has been designed to support larger antenna array sizes at both base station and UE (e.g., 256-T× 64-Rx), which enables directional transmissions of cell-specific synchronization sequence blocks (SSBs), and control and data channels, unlike LTE. Precise spatial beamforming also benefits URLLC due to the enhancement in signal-to-interference-plus-noise ratio (SINR) and corresponding drop in packet error rates. However, precise beamforming is reliant on accurate CSI, which may be more onerous to acquire in a URLLC environment with mobility, for example.

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. However, from an overall 5G system architecture perspective, improving network and UE EE is one of the basic principles of the NR design [1]. The target for infrastructure EE is a design with:

- The ability to efficiently deliver data
- The ability to provide sufficiently granular network discontinuous transmission when there is no data to transmit and network availability is maintained
- The ability to provide operator flexibility to adapt sleep durations of base stations depending on load, services, and area

For mMTC UEs, the target battery lifetime is up to 15 years. For non-mMTC UEs, the EE target is qualitative. The remainder of the article investigates these aspects for the specific case of 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 [7]. The UP latency is measured by the unidirectional time between transmission and reception of a packet at the corresponding IP layer entities of network node and UE. On the other hand, C-Plane latency is the transition time of a UE when switching from idle state to active state. In the idle state, a UE is not connected with radio resource control (RRC). After the RRC connection is set up, the UE switches from idle state into connected state and then enters into active state after moving into dedicated mode. Since the application-level throughput is dependent mainly on the UP latency, the remainder of the article dives deeper into this metric.

The total delay of a packet transmission in a cellular network can be attributed to the RAN, fronthaul, backhaul, core network, and data center or external server [11]. The total unidirectional transmission time of a 5G system can be written as

$$T = T_{Radio} + T_{Fronthaul} + T_{Backhaul} + T_{Core} + T_{Trans}$$

where

- T_{Radio} is the physical layer packet transmission time between gNB and UEs.

Precise spatial beamforming also benefits URLLC due to the enhancement in signal-to-interference-plus-noise ratio (SINR) and corresponding drop in packet error rates. However, precise beamforming is reliant on accurate channel state information, which may be more onerous to acquire in an URLLC environment with mobility, for example.

- $T_{Fronthaul}$ is the delay between the gNB RF front-end and the centralized baseband unit (if applicable).
- $T_{Backhaul}$ is the time taken to traverse the core network entities and gateways. The core network and gNB are connected by copper wires or microwave or optical fibers.
- T_{Core} is the processing time taken by the core network.
- T_{Trans} is the delay in data communication between the core network and Internet.

The component T_{Radio} includes the TTI, propagation delay, signal processing time at the receiver, and retransmission time due to packet errors. This delay is required to be less than 1 ms for URLLC, as described previously. An NR radio access network (RAN) may be connected to either an LTE evolved packet core (EPC) or a 5G next-generation core (NGC) network; reducing the delays due to $T_{Backhaul}$, T_{Core}, and T_{Trans} may not be possible if an LTE EPC is in use. Therefore, we focus on the remaining delay parameters in the sequel.

Infrastructure EE

In this section, we examine different aspects of the EE-delay trade-off from the perspective of the network infrastructure. A more rigorous evaluation of the solutions proposed in this article can be performed once the standardization of the NR RAN and system architecture has been finalized. Note that all of the solutions presented in this article are equally applicable to eMBB transmissions.

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 [12]. 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 Fig. 3 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 Fig. 3 shows an example of such a coordinated sleep schedule, where gNB A is directed to go into off mode during SFNs 40–45, and so on. In practice, the on-off switching timescale is up to the network opera-

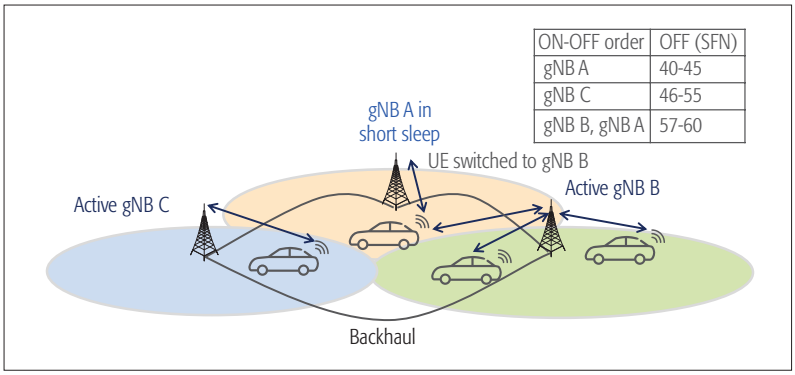

FIGURE 3. Coordinated gNB on-off switching to improve infrastructure EE without increasing delay.

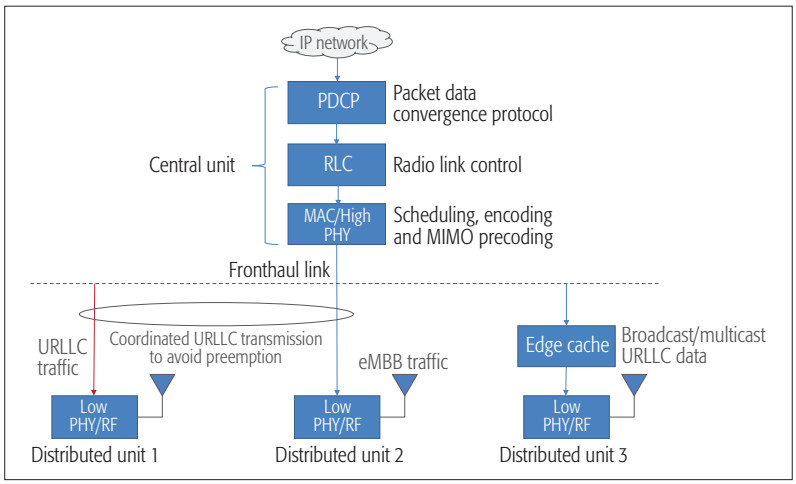

FIGURE 4. EE strategies for distributed architectures.

tor/higher-layer controller; increasing the gNB off durations can benefit network EE. By reducing the number of concurrently active gNBs, an improvement in network EE and a reduction in inter-cell interference can be achieved simultaneously. The on-off switching scheme is applied to only those UEs that have adequate SINR from the alternate gNBs, such that the same reliability is maintained when the serving gNB is switched.

Apart from inter-gNB backhaul connectivity, the main requirements for this solution are:

- Time synchronization across gNBs
- Serving gateway connectivity for UE data availability at each gNB
- Adequate transmit/receive beam quality from each substitute gNB for the UEs so as to not compromise on reliability

The first two requirements are already supported in LTE as part of inter-site coordinated multipoint (CoMP) transmission. The third requirement can be considered to be an extension of the NR mobility measurement process.

Retransmissions

Retransmission of incorrectly received data via a hybrid automatic repeat request (HARQ) protocol is a fundamental feature of LTE and NR. HARQ in LTE relies on negative acknowledgments from the receiver to proceed with a retransmission [7]. This increases the delay in packet reception, while retransmissions degrade EE in general due to multiple transmissions of the same codeword. Even though URLLC data is encoded with a very low code rate, decoding errors can occur due to unexpected interference or channel fluctuations, and retransmissions are still necessary. Furthermore, adequate resources need to be assigned to URLLC data to avoid degrading the performance of channel coding with respect to LTE due to the use of much shorter codewords.

NR features a more efficient form of retransmission where only a part of the erroneous codeword is re-sent. Specifically, multi-bit HARQ feedback from the receiver indicates which of the code block groups (CBGs) making up a codeword need to be retransmitted. While this reduces the number of bits that are retransmitted, the delay of the HARQ feedback and subsequent retransmission is still present.

A potential solution is to bundle the original transmission and retransmission in the same TTI as a preemptive measure, which can be activated after monitoring preceding HARQ feedback from UEs. Note that the preemptive retransmission approach can benefit both infrastructure EE as well as device EE. In Fig. 2, as an example, the URLLC PDSCH would therefore contain the original codeword in the first and part of the second symbol, and additional parity bits occupy the remaining portion of the second symbol. UEs that successfully decode the original codeword do not need to decode the additional parity bits. The reduction in delay is equal to the HARQ round-trip time, which is expected to be at least one slot for NR.

While the preemptive retransmission can be considered as overhead, the cost is minimized if the URLLC data is being broadcast to multiple UEs. The gNB scheduler can judiciously turn this feature on and off after analyzing the system retransmission rates. Preemptive retransmissions are also beneficial in the case of operation in unlicensed spectrum, where access to the channel is not guaranteed, and an additional channel-sensing phase would normally be required before a retransmission.

Distributed Architectures

5G systems are being designed to be amenable to centralized or cloud RAN (CRAN) architectures with a functional split between a central unit (CU) and multiple distributed units (DUs) [13]. 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, as shown in Fig. 4. 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 fronthaul links. The global control of BBU processing at the CU leads to capacity and coordination efficiencies [14], 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-decodable 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. An example is shown in Fig. 4 where 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 fronthaul 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 example is shown in Fig. 4, where an edge cache is deployed together with DU 3. A more comprehensive review of 5G caching strategies is presented in [11]. 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.

Device EE

In this section, we shift our attention to URLLC device EE. Mobile device EE is especially important since they are powered by batteries. Continuous monitoring of wideband DL control and data channels is an energy-intensive process for UEs. 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 [7]. 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. Using these features "as is" for URLLC is not desirable, as discussed next.

Discontinuous Reception and PSM

Since DRX and PSM are analogous to the on-off switching discussed in the previous section, the same trade-offs of delay vs. EE are applicable. A longer DRX duration improves UE EE, while also increasing the delay in delivering URLLC traffic since the gNB has to wait for the UE to wake up.

A potential solution would be for a single UE in a group to continuously monitor the DL and alert its DRX neighbors via device-to-device links if there are any pending DL URLLC transmissions. However, such a solution is not as attractive, since the energy costs of neighbor UE discovery and intra-UE cooperation are not easily amortized.

A better solution is to equip the URLLC UE with a secondary wake-up radio (WUR), which stays in receive mode while the primary radio system is in DRX. IEEE 802.11 is currently exploring a similar concept as part of the 802.11ba amendment for WUR [15]. The WUR hardware can be designed to incur very low power consumption, on the order of 100 μW. Whenever the gNB needs to wake up UEs in DRX, it transmits a pre-defined narrowband signal sequence to the WUR; after signal detection, the WUR triggers the main radio to immediately resume wideband RF reception. Therefore, the overall UP delay for DRX UEs can still be confined to less than 0.5 ms. An example is shown in Fig. 5, where the URLLC data delivery to a DRX UE takes up one slot in the time domain (around 59 μs) while minimizing the UE energy consumption when in sleep mode. The actual delay in data delivery will be dependent on the UE capability in terms of signal detection and time-frequency synchronization/AGC setting for the main radio after waking up.

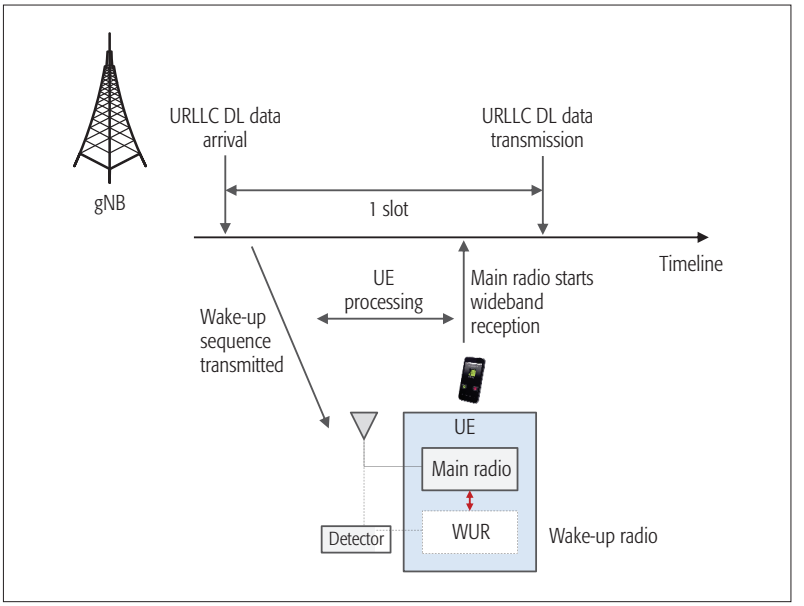

FIGURE 5. DRX UE with a wake-up radio to minimize energy expended on wideband signal reception.

Mobility Measurements

Mobility or radio resource management (RRM) measurements are an integral part of broadband systems such as LTE and NR. The RRM measurements reported from UEs are used by the network to determine if a UE should be handed over to a neighbor cell that has become better than the serving cell. Therefore, care must be taken to minimize the delays due to mobility measurements and handover. The general RRM process entails the UE scanning one or more frequencies in order to detect synchronization sequences of neighbor cells, followed by measuring the signal strength of associated reference signals [6]. Compared to LTE, the energy cost can be more prohibitive in NR since SSBs are transmitted less frequently (every 5 ms, 10 ms, or more, instead of every 1 ms), can have a variable location in the frequency domain, and can be transmitted with different beamforming directions. Therefore, URLLC UEs will expend more energy if they blindly scan a carrier to detect SSBs.

A candidate solution to improve both EE and latency of mobility measurements is depicted

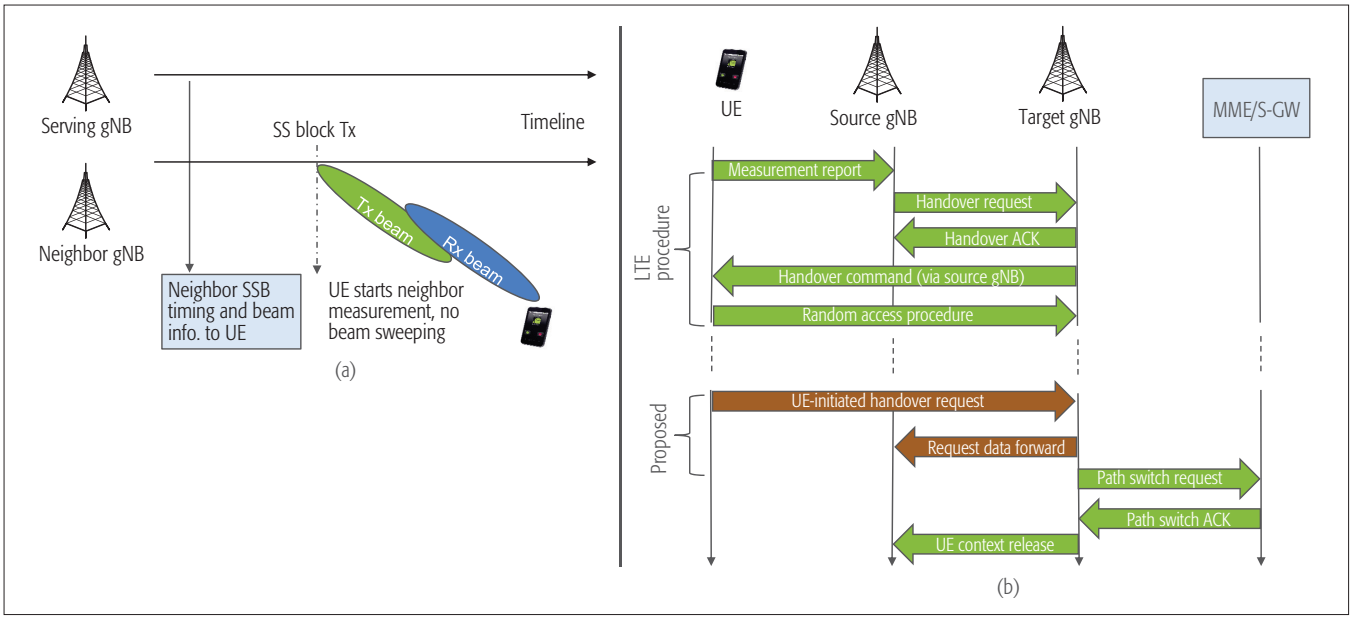

FIGURE 5. a) Time-frequency and beam assistance information provided to UE for neighbor gNB measurements; b) UE-initiated handover request to reduce handover delay.

in Fig. 6a. Here, the serving gNB assists UEs by providing the time-frequency location and beamforming information of neighbor gNB SSBs. UE EE is improved since the UE turns on its radio and performs measurements at the SSB transmission time, without the need for a blind search. The a priori knowledge of the SSB beam also helps the UE avoid a costly beam-sweeping procedure. The assistance information can be included together with the rest of the (large) radio resource configuration (RRC) information sent to UEs, thereby reducing the additional energy consumption needed to receive it. The provision of assistance information by the serving gNB for its own SSBs is currently under discussion in 3GPP; the extension to neighbor information would be a beneficial addition.

Handover Procedure

The delay of the handover procedure in LTE can be as large as several hundred milliseconds. As shown in Fig. 6b, it consists of the following phases when an X2 interface is present between base stations (eNBs).

Reporting: UE measurement reports are sent to the source eNB when triggered by a pre-defined event. The source eNB decides to initiate HO (approximately 80–400 ms due to time-averaging of layer 1 measurements).

Preparation: The source eNB sends a handover (HO) request to the target eNB, which is either accepted or rejected based on the target's call admission control (approximately 20 ms).

Execution: The source eNB forwards the HO command message from the target eNB to the UE. The data path is switched to the target eNB by the core network (mobility management entity, MME, and serving gateway). The UE performs a random access procedure to acquire UL synchronization with the target eNB before starting data reception/transmission. Data transfers are interrupted during this phase (approximately 50 ms).

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

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.

The UE needs to be aware of pre-defined UL resources on the target gNB to send the request; this information can be either obtained by reading the system information broadcast by the target gNB, or received as configuration from the source. The UE also alerts its serving gNB once the target gNB has responded, so as to avoid unnecessary resource allocations from the source gNB. Therefore, both delay and EE metrics for the UEs can be improved due to faster HO to a better gNB; the total handover delay can be reduced by about 50 percent compared to exist-

ing HO. The reduction in network control of the HO process is the reason why such a mechanism has not yet been implemented in cellular systems; however, the unique delay constraints of URLLC make such solutions more compelling.

Potential Research Issues

This work 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. NR has introduced a new UE state known as INACTIVE where it remains connected to the network; this falls between the existing IDLE and CONNECTED states of LTE — the EE performance of the INACTIVE state remains open. Apart from over-the-air signaling, numerous non-access stratum procedures such as UE authentication and security measures contribute to delays before UEs successfully attach to the network and begin receiving data. A holistic treatment of all such delays is needed from a system perspective.

Another area of interest is a detailed study of URLLC EE/delay in unlicensed spectrum such as the 60 GHz band. In unlicensed spectrum, regulations require continuous sensing of the channel to obtain transmission opportunities and impose a limit on the maximum transmission time. Such constraints hamper both EE and delay performance, and call for new solutions in the case of URLLC. 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.

Apart from over-the-air signaling, numerous non-access stratum procedures such as UE authentication and security measures contribute to delays before UEs successfully attach to the network and begin receiving data. A holistic treatment of all such delays is needed from a system perspective.

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 Pro and 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-Fall 2016*, 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, July 2017, pp. 62–68.
[15] IEEE 802.11-16/1045r9, "A PAR Proposal for Wake-Up Radio," 2016.

Biography

Amitav Mukherjee [S'06, M'13, SM'17] received his B.S. degree from the University of Kansas, Lawrence, in 2005, his M.S. degree from Wichita State University, Kansas,, in 2007, both in electrical engineering, and his Ph.D. degree in electrical and computer engineering from the University of California, Irvine in 2012. He is currently a Distinguished Member of Technical Staff at Verizon, where he is a technical architect for next-generation LTE-A and 5G radio access networks. He previously was a senior researcher and standards delegate at Ericsson Research, San Jose, California, from 2014 to 2017. From 2012 to 2014, he was a wireless systems researcher at Hitachi America Ltd., Santa Clara, California. His research interests encompass statistical signal processing and wireless communications, with over 70 publications and 80 pending/issued patents in these areas.