# Energy-Efficient Resource Allocation for Ultra-Reliable Wireless Communication

Deepa K.S[1], S.PAudline beena[2], D.Rajinigirinath[3]
[1]*PG Student, Sri Muthukumaran Institute of Technology, Chennai, Tamil Nadu, India*
[2]*Assistant professor, Sri Muthukumaran Institute of Technology, Chennai, Tamil Nadu, India*
[3]*Head of the Department, Sri Muthukumaran Institute of Technology, Chennai, Tamil Nadu, India*

***Abstract*- Ensuring ultra-reliable and low-latency communication (URLLC) for 5G wireless networks and beyond is of capital importance and is currently receiving tremendous attention in academia and industry. At its core, URLLC mandates a departure from expected utility-based network design approaches, in which relying on average quantities. Instead, a principled and scalable framework which takes into account delay, reliability, packet size, network architecture, and topology (across access, edge, and core) and decision making under uncertainty is sorely lacking. The overarching goal of this article is a first step to fill this void. Towards this vision, after providing definitions of latency and reliability, we closely examine various enablers of URLLC and their inherent tradeoffs. Subsequently, we focus our attention on a plethora of techniques and methodologies pertaining to the requirements of ultra-reliable and low-latency communication, as well as their applications through selected use cases. These results provide crisp insights for the design of lowlatency and high-reliable wireless networks.**

**Index Terms-Ultra-reliable low-latency communication, 5G and beyond, resource optimization, mobile edge computing.**

## 1. INTRODUCTION

The phenomenal growth of data traffic spurred by the internet-of-things (IoT) applications ranging from machine-type communications (MTC) to mission-critical communications (autonomous driving, drones and augmented/virtual reality) are posing unprecedented challenges in terms of capacity, latency, reliability, and scalability. This is further exacerbated by: i) a growing network size and increasing interactions between nodes; ii) a high level of uncertainty due to random change s in the topology; and iii) a heterogeneity across applications, networks and devices.

The stringent requirements of these new applications warrant a paradigm shift from reactive and centralized networks towards massive, low-latency, ultra-reliable and and error rates for various connectivity protocols is proactive 5G networks. Up until now, human-centric communication networks have been engineered with a focus on improving network capacity with little attention to latency or reliability, while assuming few users. Achieving ultra-reliable and low-latency communication (URLLC) represents one of the major challenges facing 5G networks. URLLC introduces a plethora of challenges in terms of system design. While enhanced mobile broadband (eMBB) aims at high spectral efficiency, it can also rely on hybrid automatic repeat request (HARQ) retransmissions to achieve high reliability. This is, however, not the case for URLLC due to the hard latency constraints. Moreover, while ensuring URLLC at a link level in controlled environments is relatively easy, doing it at a network level and over a wide area and in remote scenarios (e.g., remote surgery) is notoriously difficult. This is due to the fact that for local area use cases latency is mainly due to the wireless media access, whereas wide area scenarios suffer from latency due to intermediate nodes/paths, fronthaul/backhaul and the core/cloud. HARQ). By contrast, the performance requirements of URLLC are more stringent with a target BLER of depending on the use case. From a physical-layer perspective, the URLLC design is challenging as it ought to satisfy two conflicting requirements: low latency and ultra-high reliability

On the one hand, minimizing latency mandates the use of short packets which in turns causes a severe degradation in channel coding gain. On the other hand, ensuring reliability requires more resources

(e.g., parity, redundancy, and re-transmissions) increasing latency (notably for time-domain redundancy). Furthermore, URLLC warrants a system design tailored to the unique requirements of different verticals for which the outage capacity is of interest (as opposed to the Ergodic capacity considered in 4G). This ranges from users (including cell edge users) connected to the radio access network which must receive equal grade of service, to vehicles reliably transmitting their safety messages and industrial plants whereby sensors, actuators and controllers communicate within very short cycles.

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). 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. 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.

## 2. PROPOSED SYSTEM

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.

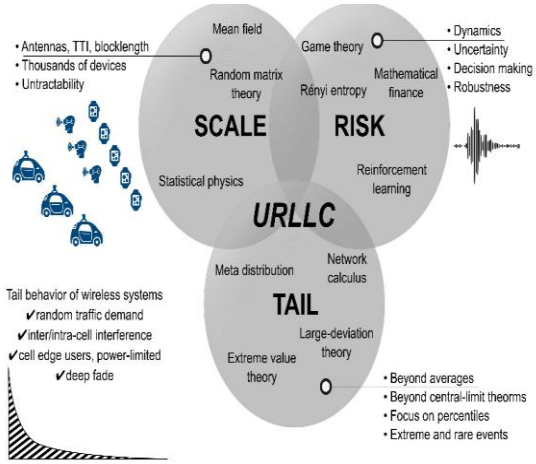


Fig 1. Anatomy of the URLLC building blocks, composed of tail, scale and risk alongside their unique characteristics.

A. Low-Latency

A latency breakdown yields deterministic and random components that are either fixed or scale with the number of nodes. While the deterministic component defines the minimum latency, the random components impact the latency distribution and more specifically its tails. Deterministic latency components consist of 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.

B. Reliability

The main factors affecting reliability stem from: i) collisions with other users due to uncoordinated channel access; ii) coexistence with other systems in the same frequency bands; iii) interference from users in adjacent channels; iv) Doppler shifts from moving devices, v) difficulty of synchronization, outdated channel state information, time-varying channel effects or delayed packet reception. Reliability at the physical layerlevel (typically expressed in block error rate) depends on factors such as the channel, constellation, error detection codes, modulation technique, diversity, retransmission mechanisms, etc. A variety of techniques to increase reliability include using low-rate codes to have enough redundancy in poor channel conditions, retransmissions for error correction, and ARQ at the transport layer.

While reinforcement learning aims at maximizing the expected utility of an agent (i.e., a transmitting node), risk-sensitive learning is based on the fact that the utility is modified so as to incorporate the risk (e.g., variance, skewness, and other higher order statistics).

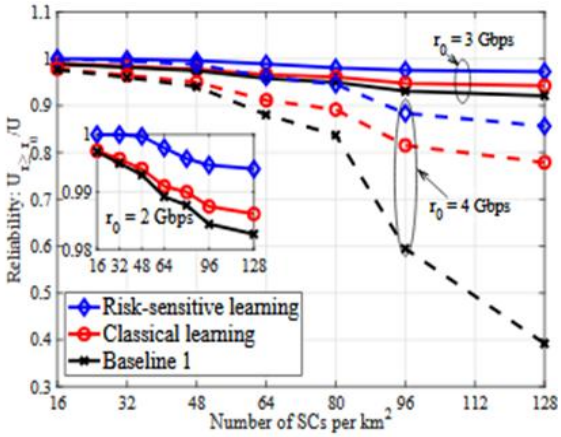


Fig 2: The trade-off between reliability and network density

3. SYSTEM ARCHITECTURE

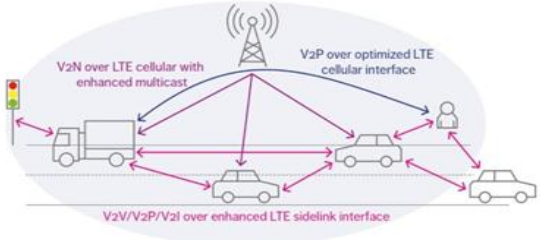


Fig 3: System Architecture

5G systems are being designed to be amenable to centralized or cloud RAN (CRAN) architectures with a functional split between a central unit (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.

To minimize the VR service latency, players offload their computing tasks which consist of rendering high definition video frames to the edge servers over mmWave links. First, players send their tracking data, consisting of their poses (location and rotation coordinates) and game play data, in the uplink to an edge server. The edge server renders the corresponding player's frame and transmits it in the downlink. Since edge servers are typically equipped with high computation power graphical processing units (GPUs), computing latency is minimized as compared tolocal computing in the player's HMD. In addition to minimizing computing latency, reliable and low latency communication is needed to minimize the over-the-air communication latency.

Ensuring a reliable link in mmwave-enabled VR environment is a daunting task since the mmWave signal experiences high level of variability and blockage. Therefore, we investigate MC as an enabler for reliable communication, in which a gaming arcade with 8x8 game pods, served by multiple mmwave access points connected to edge servers is assumed. We model the user association to edge servers as a dynamic matching problem to minimize service latency such that users with a link quality below a predefined threshold are served via MC.

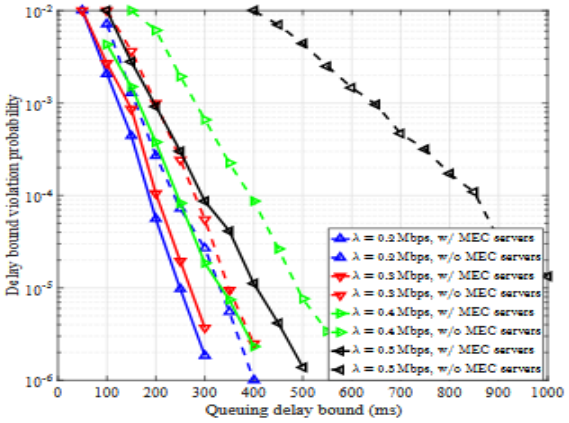


Fig 4: Delay bound violation probability versus queuing delay bound.

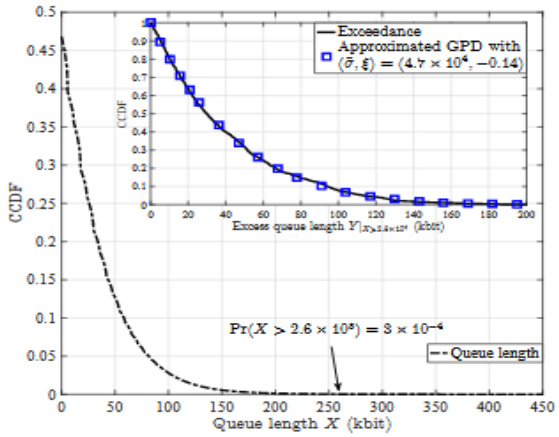


Fig 5: Tail distributions of a given UE's task queue length, queue length exceedance over threshold, and the approximated GPD of exceedances

## 5. CONCLUSION

Enabling URLLC warrants a major departure from average-based performance towards a clean-slate design centered on tail, risk and scale. This has reviewed recent advances in low-latency and ultra-high reliability in which key enablers have been closely examined. Several methodologies stemming from adjacent disciplines and tailored to the unique characteristics of URLLC have been described. In addition, via selected use cases we have demonstrated how these tools provide a principled and clean-slate framework for modeling and optimizing URLLC-centric problems at the network level. This article will help foster more research in

URLLC whose importance will be adamant in beyond 5G and 6G networks, with the slew of unforeseen applications.

REFRENCES


[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.