SCIENCE CHINA
Information Sciences

February 2021, Vol. 64 120301:1–120301:12
https://doi.org/10.1007/s11432-020-2852-1

• REVIEW •
Special Focus on Ultra-Reliable Low-Latency Communications in Wireless Networks

# Ultra-reliable and low-latency communications: applications, opportunities and challenges

Daquan FENG[1], Lifeng LAI[1], Jingjing LUO[2], Yi ZHONG[3*], Canjian ZHENG[2] & Kai YING[4]

[1]*College of Electronic and Information Engineering, Shenzhen University, Shenzhen* 518060*, China;*
[2]*School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen* 518055*, China;*
[3]*School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[4]*Sharp Laboratories of America, Camas* 98607*, USA*

Received 15 January 2020/Accepted 24 March 2020/Published online 20 January 2021

**Abstract** In the upcoming 5G and beyond systems, ultra-reliable and low latency communication (URLLC) has been considered as the key enabler to support diverse mission-critical services, such as industrial automation, remote healthcare, and intelligent transportation. However, the two stringent requirements of URLLC: extremely low latency and ultra-strict reliability have posed great challenges in system designing. In this article, the basic concepts and the potential applications of URLLC are first introduced. Then, the state-of-the-art research of URLLC in the physical layer, link layer and the network layer are overviewed. In addition, some potential research topics and challenges are also identified.

**Keywords** ultra-reliable and low-latency, advanced transmission technologies, physical layer, network design, real-time guarantee

**Citation** Feng D Q, Lai L F, Luo J J, et al. Ultra-reliable and low-latency communications: applications, opportunities and challenges. Sci China Inf Sci, 2021, 64(2): 120301, https://doi.org/10.1007/s11432-020-2852-1

## 1 Introduction

The emerging fifth-generation (5G) of wireless cellular networks has been considered as a revolutionary technology that will reshape many different industries and also improve the living standards for people in the planet as well as bring huge economic benefits in the coming years. According to a recent market analysis report, 5G is expected to enable 13.2 trillion dollars of global economic output in 2035 [1]. Therefore, the world's leading operators and manufactures are striving to develop the 5G technology. The main focus of the 5G networks is to increase the capacity, connection density, and also network availability while enhancing reliability and reducing latency [2]. To separate the diverse performance requirements and support a variety of services, 5G has considered three major service categories: enhanced mobile broadband (eMBB) to provide high data rate and spectral efficiency, massive machine-type communication (mMTC) enabling the access of a large number of machine-type devices, and ultra-reliable and low latency communication (URLLC) to support the mission-critical communications [3,4]. Among the three service categories, URLLC could be the most challenging one to achieve. The other two services, i.e., eMBB and mMTC, can evolve based on the traditional design of cellular networks, to provide a larger rate for a larger number of equipment. URLLC brings new requirements, i.e., a hard delay constraint and a high reliability. Thus, URLLC has attracted increasing attention from both the academic and industrial communities [5, 6].

URLLC is crucial for enabling mission-critical services and has the potential to transform many different industries, as illustrated in Figure 1.

The potential applications include: industrial automation [7], remote healthcare [8], intelligent transportation [9], augmented reality/virtual reality (AR/VR) [10] and smart grid [11].

* Corresponding author (email: yzhong@hust.edu.cn)

 

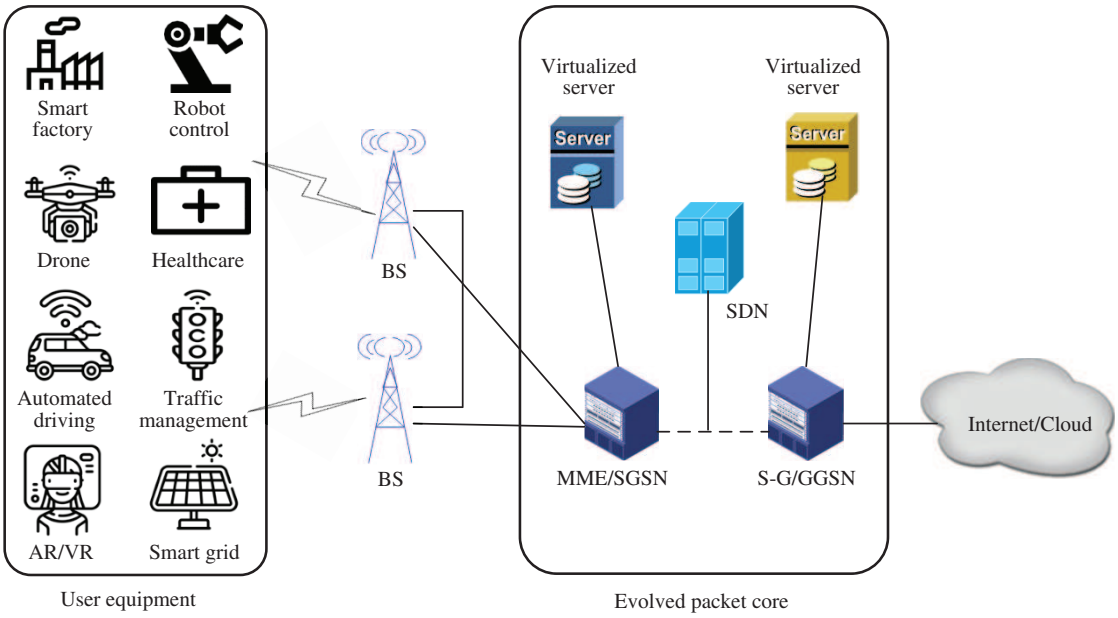


**Figure 1** (Color online) Applications and network architecture of URLLC.

**Table 1** The requirements of mission-critical services in URLLC

| Mission-critical services | E2E latency (ms) | Reliability (%) |
|---|---|---|
| Industrial automation | 0.25–10 | 99.9999999 |
| Remote healthcare | $< 30$ | 99.999 |
| Intelligent transportation | 1–100 | 99.9999 |
| AR/VR | 0.4–20 | 99.999 |
| Smart grid | 3–20 | 99.999 |

• Industrial automation. As one of the key enabling technologies in the fourth industrial revolution, URLLC is expected to deploy in factories to automate the mission-critical processes such as robot motion control and tactile iteration. Compared the traditional wired connections, using URLLC technology can help to reduce the operational cost and also improve the deployment flexibility.

• Remote healthcare. In the telemedicine services such as remote diagnosis, remote surgery, and emergency response, the date transmission is life critical. URLLC supports the timely and reliably delivery of various data, and thus it enables to provide better medical services to people in the remote areas and also reduce regional imbalance in the health workforce.

• Intelligent transportation. URLLC allows the exchange of information among vehicles, infrastructures, and pedestrians with high reliabilities in a very short time. Thus it can help to enhance road safety and improve traffic efficiency. The typical usage cases include: automated overtake, cooperative collision avoidance, high density platooning, and traffic management.

• AR/VR. AR aims to enhance the real environment with graphics whereas VR focuses on creating a totally virtual and simulated experience. Both of them have the extremely strict requirements of the latency and reliability. URLLC supports the reliable real-time data transmission for AR/VR, and could solve the cyber sickness and the uncomfortable user experience caused by lengthy delay between action and response. Thus, URLLC is helpful to accelerate the prosperity of AR/VR industry.

• Smart grid. Smart grid consists of many mission-critical services, such as fault diagnosis and accurate positioning, fault isolation and system restoration as well as remote decoupling protections. Because there are a large number of power distribution grid lines and stations, the cable/fibre based solution will bring a huge deployment cost. Therefore, URLLC can provide a cost-effective alternative solution for fibre optics to carry out the real-time protection and control over the distributed grid lines and stations.

In addition, the specifical requirements in terms of latency and reliability of aforementioned applications are shown in Table 1.

Based on the potential use cases and applications, the 3rd generation partnership project (3GPP) has specified some service requirements and system features for general URLLC. Specifically, the reliability

**Table 2** 4G-LTE vs. URLLC

| | 4G-LTE | URLLC in 5G and beyond |
|---|---|---|
| BLER | $10^{-1}$ | $10^{-5} - 10^{-9}$ |
| End-to-end latency | 30–100 ms | Several milliseconds |
| User plane latency | 4 ms range | At most 1 ms |
| Packet sizes | $\gg$ 100 bytes | Tens to hundreds of bytes |
| TTI | 1 ms | No more than 0.2–0.25 ms |

of 99.999%, i.e., block error rate (BLER) of $10^{-5}$, should be guaranteed for a single transmission of a 32-byte long packet within 1 ms. Thus, the target user plane latency of 0.5 ms should be supported for both uplink (UL) and downlink (DL) [12]. In addition, to enable a variety of applications with a variety of latency and reliability constraints, a target BLER of $[10^{-5} - 10^{-9}]$ together with the end-to-end (E2E) latency of several milliseconds has also been suggested to be supported in URLLC [13].

To meet the stringent requirements, a shorter transmission time interval (TTI), such as no more than 0.2–0.25 ms, and a smaller packet size, such as tens to hundreds of bytes, are critical to reduce the coding/decoding delay and transmission delay as well as the processing delay. However, the current fourth-generation long-term evolution (4G-LTE) system is unable to satisfy all the performance requirements for URLLC. In 4G-LTE system, the typical BLER is $10^{-1}$ while the E2E latency is in the range of 30–100 ms. Furthermore, 3GPP has specified the minimum user-plane latency of 4 ms and the TTI of 1 ms for 4G-LTE systems. In addition, the packet size is much longer than hundreds of bytes in the 4G-LTE system which mainly focuses on achieving high throughput. A brief comparison between 4G-LTE and URLLC is shown in Table 2.

To achieve the goal of URLLC and meet the requirements of diversified traffic in a large wireless network [14, 15], the network architecture needs to be reconsidered. Ref. [16] provided an up-to-date comprehensive survey of the IEEE TSN and IETF DetNet standards and related researches. The latency of a packet in a wireless network is contributed by wireless access, E2E delay, backhaul, core network. The total one way transmission delay can be written as

$$T = T_q + T_t + T_d + T_b. \tag{1}$$

- $T_q$ is the queuing delay which depends on the arrival rate of packets in the queue and the priority level of the users.
- $T_t$ is the transmission time of the packet which depends on the channel condition, payload size, frame structure and the transmission strategy.
- $T_d$ is the decoding delay, which cannot be ignored when the latency is very low.
- $T_b$ is the backhaul delay between the BS and the core network. It depends on the transmission material of the wired connections.

The E2E delay is approximately equal to $2T$, and Ref. [17] investigated how the delay accumulates from PHY to transport layer, and showed how to characterize the E2E delay into several components.

Many approaches such as physical layer design (e.g., channel coding, frame structures, HARQ) and advanced transmission technology (e.g., grant-free transmission, device-to-device (D2D) communications, massive multiple-input and multiple-output (MIMO) and unmanned aerial vehicle (UAV) communications) as well as network layer optimization have been proposed to meet the stringent requirements in URLLC. However, there remain many technical challenges to be addressed. In this article, we will overview the potential implementation methods of URLLC from both the aspect of the physical layer and the aspect of network architecture. In addition, we also identify the future research challenges. The rest of this article is organized as follows. In Section 2, the physical design for URRLC is introduced. Then, link layer transmission technologies are presented in Section 3. In Section 4, the implementation of URLLC in the large-scale wireless networks is addressed. Finally, the research challenges are discussed in Section 5.

## 2 Physical layer design for URLLC

The target of physical layer processing is to deliver protocol data unit (PDU) from MAC layer successfully. The PDU is also named as transport block (TB) in physical layer. As mentioned in Section 1, for URLLC

in 5G NR, the target user plane latency is 0.5 ms for both UL and DL while the target reliability is $10^{-5}$ for 32-byte long packet transmission within 1 ms. In general, factors impacting latency in physical layer can be categorized as waiting time, transmission time, user equipment (UE) processing time and NR Node B (gNB) processing timing. The waiting time and transmission time can be shorten by scheduling and frame structure design, while UE/gNB processing time is highly related to device capability, which is up to implementation. On the other hand, reliability in physical layer depends on channel coding, hybrid automatic repeat request (HARQ) as well as scheduling.

### 2.1 Flexible frame structure

In physical layer, waiting time includes delay waiting for scheduling request (SR) opportunity, delay between a UL grant and corresponding UL data transmission, delay between a DL assignment and corresponding DL data transmission, and delay between a DL data reception and corresponding acknowledgement (e.g., acknowledgement (ACK)/not-acknowledged (NACK)) transmission on the UL. Additionally, waiting time is in unit of TTI which also serves as the minimum scheduling unit. For example, in 4G-LTE, the minimum periodicity of SR opportunities is 1 TTI and the delay between an UL grant and corresponding UL data transmission is 4 TTIs. In 4G-LTE, TTI is presented by a subframe, which consists of 14 orthogonal frequency-division multiplexing (OFDM) symbols with normal cyclic prefix (CP). Given that the subcarrier spacing (SCS) is 15 kHz, the duration of a subframe/TTI is 1 ms. In this case, when a TB arrives, the average waiting time for a UE to transmit SR is 0.5 ms (0.5 SR period). After receiving a UL grant from gNB, the waiting time for a UE to transmit the corresponding UL data is 4 ms.

Therefore, we can conclude that TTI should be shorten so that the waiting time can be reduced to meet the latency requirement of URLLC. The other motivation to shorten the TTI is that long transmission time for URLLC should be avoided because the TB size of URLLC data is usually very small. In general, there are two methods to shorten the TTI by adjusting the frame structure. One is frequency domain method which increases the SCS to decrease the symbol duration. The other is the time domain method which directly reduces the number of OFDM symbols in a TTI.

### 2.2 Flexible scheduling timeline

In 4G-LTE, the waiting time is fixed by specification. In addition, the waiting time is set long enough so that even UE with limited capability can also support the HARQ/scheduling timeline. However, the low latency may not be achieved. In NR, flexible waiting time is introduced to enable dynamic scheduling and to meet different latency/capability requirements.

Here, let $K_0$, $K_1$, and $K_2$ slots denote the delay between a DL assignment and corresponding DL data transmission, the delay between a DL data reception and corresponding acknowledgement transmission on the UL, and the delay between a UL grant and corresponding UL data transmission, respectively. A set of values for $K_0$ ($K_1$, or $K_2$) is initially configured by the higher layer. Then, the physical layer selects one value from the set for the scheduling timeline. The value can be 0, which means the waiting time is within a single slot.

Generally speaking, to achieve the low latency, the waiting time should be always set as the minimum value. However, the minimum waiting time that a UE/gNB can support depends on the UE/gNB processing time. Because gNB is powerful, $K_0$ can be always set as 0 for URLLC transmission. For the value of $K_1/K_2$, we should take in account the processing time that UE consumes to demodulate/decode a UL grant or a DL transmission. Quick processing of a control message or a data packet may not be available for some UEs with limited capability, so that $K_1/K_2$ cannot be 0 for these UEs. It is believed that URLLC UE should be always enabled with the highest capability and the minimum value from the configured set for waiting time can be selected.

### 2.3 Enhanced HARQ feedback

In the current hybrid automatic repeat request- acknowledgement (HARQ-ACK) transmission, the transmitter needs to retransmit the packets for avoiding packets error and improving transmission reliability. This requires the receiver to issue feedback to the transmitter for each received packet in control channels. However, more delay components, such as the feedback round-trip-time (RTT), may be introduced in such HARQ-ACK transmission scheme and significantly affect the latency performance. To reduce the feedback RTT and improve reliability, 3GPP release 16 has considered two enhanced HARQ feedback

schemes. One is enabling more than one physical uplink control channel (PUCCH) for HARQ-ACK transmission within a slot and the other one is the enhanced reporting procedure/feedback for HARQ-ACK. In particular, enabling more than one PUCCH for HARQ-ACK transmission within a slot can provide the fast HARQ-ACK feedback to reduce the latency and support the separated HARQ-ACK feedbacks for URLLC and eMBB. Enhanced reporting procedure/feedback is expected to support enhanced HARQ-ACK multiplexing on PUSCH and PUCCH, where at least two HARQ-ACK codebooks can be simultaneously constructed to support the different service types for a UE.

### 2.4 DL pre-emption

Owing to the sporadic nature of URLLC service, it is not able to predict when URLLC data will arrive. If the slot is already allocated to a long transmission like eMBB traffic, the upcoming URLLC transmission will be delayed until next available slot. The delay can be as long as 1 ms, so that the latency requirement may not be satisfied.

In NR, an aggressive solution is agreed that URLLC data can pre-empt the resource of the ongoing eMBB transmission. In other words, URLLC data may occupy any part in the frequency band and any OFDM symbol of the eMBB transmission if gNB tries to schedule the URLLC traffic as early as possible.

In this case, the latency requirement of DL URLLC transmission can always be met with the sacrifice of eMBB transmission performance. Some solutions to protect eMBB or minimize the impact of URLLC pre-emption to eMBB have been discussed in the state of the art, but the discussion is out of scope in this article.

### 2.5 Channel coding, CQI and MCS table for URLLC

In 5G NR, low-density parity-check (LDPC) code is used for data channel while polar code is adopted for control channel. To guarantee the reliability of data transmission, first of all, a BLER target should be defined. Then, according to the channel quality indication (CQI) from channel estimation, an appropriate modulation and coding scheme (MCS) will be selected from a look-up table to meet the BLER target. Generally speaking, for the extremely low BLER target, high modulation order like 256QAM should be avoided and coding rate should be low enough. The accuracy of channel estimation is also a big challenge.

## 3 Advanced transmission technologies

Advanced transmission technologies such as grant-free (GF) random access, D2D communications, massive MIMO and UAV communications are among the key enablers to support URLLC services. In this section, we overview the current progress of these technologies and introduce them one by one.

### 3.1 Grant-free random access

In current 4G-LTE system, UEs follows grant-based scheme to guarantee the reliable transmission in UL. In particular, the grant-based scheme involves a four-step handshake process between the UEs and base stations (BSs) to avoid any potential collision. However, it comes at the expense of lengthy latency for the access grant response. For the tight latency requirements of the mission-critical applications in URLLC, such a grant-based scheme is not applicable. Therefore, the GF random access, which allows UEs to eliminate the phases of the handshake process and get fast access, is an attractive option to minimize the latency of UL access.

The general idea of GF transmission is to preconfigure/allocate periodic dedicated resources for URLLC service. It works well for the periodic traffic. However, owing to the sporadic nature of URLLC traffic, UE can skip the GF resource if there is no data to be transmitted. If the GF resource is dedicated to a UE for URLLC service and cannot be used by other UEs or services, then the resource may be wasted. To better utilize the resource, one possible solution is to allow different UEs and different services sharing the GF resource. However, it may cause collisions occurred between the UEs which will degrade the transmission reliability and increase transmission latency. In [18], the authors have proposed a contention-based transmission scheme for the sporadic URLLC traffic. To improve the reliability, the diversity transmission is adopted. In addition, the authors in [18] have derived the optimal number of repetitions that satisfy the strict reliability and latency constraints. In [19], the combination of non-orthogonal multiple access (NOMA) and GF transmission has been considered, where NOMA techniques

are used in the transmitter side to improve resource efficiency and the advanced receivers is adopted in receiver side to minimize the collisions.

### 3.2 D2D communication

D2D communication can support the direct communication between the nearby mobile devices without going through the cellular BSs or access points [20]. Therefore, it can significantly reduce the E2E delay. In addition, the additional D2D links can increase the communication freedom of D2D users [21] and thus help to improve the reliability, especially when the D2D is at the cell edge or in the deep fading state. In [22], D2D communication has been considered to provide the URLLC services for the machine-type communications, such as vehicle-to-vehicle communications, factory automation and remote control.

To improve the transmission reliability between a controller and a group of actuators in a typical wireless industrial automation scenarios, the authors in [23] have proposed a novel D2D-based two-phase transmission scheme. The results in [23] show that with inter-cell interference, the proposed scheme can achieve a probability of URLLC above 99.99% and much better than the existing URLLC schemes.

### 3.3 Massive MIMO

With a large number of antennas, massive MIMO systems are able to overcome the impact of channel fading by the spatial diversity, which makes it a key enabling technique for the URLLC. In addition, owing to multiple antenna systems can provide high SINR and also spatial multiplexing capability. It can contribute to support the services with stringent reliability and latency requirements, such as the tactile internet. In [24], the influence of the large antenna arrays at the receiver side to achieve the URLLC revised has been studied. The results have shown that with a massive MIMO BS, even the single antenna transmitter can satisfy the ultra-high reliability requirement. In [25], with the constraints on reliability and latency, the millimeter wave-enabled massive MIMO networks have been studied. By adopting the Lyapunov technique, it is shown that 99.99% reliability can be guaranteed and the latency can also be significantly reduced. Considering the acquisition of accurate instantaneous CSI in the multi-antenna system is difficult, the authors in [26] have proposed a linear minimum error probability detector to achieve URLLC under the imperfect CSI conditions.

### 3.4 UAV communication

UAV communication is another promising solution that supports URLLC services because it can provide the ubiquitous radio coverage and the better link qualities by high possibility of line-of-sight (LoS) communication links. Owing to the flexible and dynamic features of UAVs as flying relays to support URLLC services, it brings some new challenges for the UAV communication, such as limited energy and caching and communication resource. In [27], a resource allocation optimization scheme has been proposed to enhance the efficient resource utilization for UAVs. In addition, with the latency constraint, the finite blocklength regime and the UAV's location have been jointly considered in [28] to minimize the decoding error probability.

Because the effect of distance attenuation and shadowing, the quality of service (QoS) and the network availability requirements of URLLC in terrestrial wireless networks are hard to be guaranteed. In [29], UAV communication has been considered to provide the QoS and the network availability requirements of URLLC, where a joint altitude of UAV and bandwidth allocation optimization problem has been proposed to minimize the total bandwidth of URLLC. It has shown that multiple UAV communication links should be established to satisfy the requirement of URLLC in the urban areas.

## 4 Network design for URLLC

### 4.1 Latency and reliability

5G has three application scenarios, namely eMBB, URLLC and mMTC. Different application scenarios also have different requirements for delay. For example, in URLLC the corresponding E2E latency as low as 1 ms needs to be met with reliability as high as 99.999%. In the scenario of mMTC, there are massive devices that need to be connected with low delay at low power. Meeting diverse delay requirements is one of the most critical goals for the design of 5G wireless networks. In order to explore the impact of

scheduling policy on delay under different services, Ref. [30] proposed the notion of delay outage and evaluated the effect of different scheduling policies on the delay performance. There are many tradeoff issues [31] about latency that need to be considered, for the network metrics such as capacity, latency, reliability, power and security, once you optimize one of them, this may result in degradation of another metric. In the following, we discuss existing approaches for reducing delay.

• Multi-connectivity. Because the 1 ms delay required by 5G is less than the channel coherence time, at the same time, there are path loss and shadowing, retransmission can hardly improve the transmission reliability. In order to ensure the reliability as the same as low latency, the concept of multi-connectivity was proposed. In [32], by using both D2D and cellular links to transmit each packet and exploiting multi-connectivity, a framework to maximize the available range was established, which was defined as the maximal communication distance subject to the network availability requirement. Ref. [33] tried to provide a definition for multi-connectivity, and the authors identified main scheduling categories, compared different network architectures and considered different layers for implementing multi-connectivity.

• Mobile edge computing (MEC). MEC is one of the key technologies to achieve the low latency for 5G, enabling the wireless network to be transformed into a multi-functional service platform for various industrial applications. MEC concentrates computing resources and services at the network edge nodes, providing the URLLC service to users and reducing the network costs. MEC is also necessary to support new human-centered services and matter-centered internet of everything applications.

• Cache network. Because the latency of data transfer also depends on the latency between the core network and the BS, caching networks can be used to reduce latency by storing popular data at the edge of the network. The large latency may be caused by too many user requests in peak-traffic hours, and caching in information-centric networks can be used as one of the most promising candidates to decrease the large latency. There are two main approaches of caching: centralized and distributed. The research on caching mainly focuses on the trade-off between storage and latency [34, 35], content assignment and delivery policy scheme, and cache placement approach. Ref. [36] considered a novel caching framework with erasure code called functional caching, based on the arrival rates of different files, placement of file chunks on the servers, and the service time distribution of storage servers, and an optimal functional caching placement and the access probabilities of the file request from different disks are considered.

Although there have been many existing studies on ultra-low latency, there are still several open issues and challenges for future research. New modulation and coding schemes, massive MIMO, optimization of radio resource allocation, carrier aggregation in millimeter wave, interference problems in high-density industrial automation scenarios and priority of data transmission still need to be addressed. In indoor and outdoor environments that are still evolving, channel modeling with low latency, path loss, NLOS beamforming, angular propagation and the mobility [37] needs to be studied. In addition, the research on the latency problem is by no means independent. It is closely related to other network performance such as network capacity, reliability, and energy efficiency.

### 4.2 Real-time guarantee

With the rapid popularity of handheld devices and ubiquitous connections, real-time status updates have become an important and ubiquitous form of communication. Traffic monitoring center needs to master the speed of the car on the road, the surrounding environment, traffic conditions, position and tire pressure, etc. to warn the driver in a timely manner preventing accidents. Detectors installed in the forest can prevent fires in a timely manner based on the concentration of smoke, air temperature and humidity and you can find your friends dating with you by mobile phone or smart watch. These scenarios share a common description: the monitor wants to get information as timely as possible from the remote of interest. Therefore, research on the real-time update system is more important in order to obtain the freshest information.

Indicators such as "delay" and "throughput" can no longer accurately measure the freshness of information. In 2010, Kaul et al. [38] proposed a new concept age of information (AoI) to evaluate the performance of a real-time status update system. At time $t$, the monitor receives a packet with a time stamp $u(t)$, which is the generation time of the packet at the source, and then $t - u(t)$ is defined as AoI. AoI is not only affected by transmission delay, but also related to the update rate from the source. And it is a concept that can measure real-time more accurately and comprehensively. In recent years, many scholars have conducted a lot of research based on AoI [39].

At first, some studies analyzed AoI based on different rules and queues and tried to optimize it. These included single-source single-server queues [38], the $M/M/1$ LCFS queue with pre-emption in service [40], and the $M/M/1$ FCFS system with multiple sources [41]. Because of these initial efforts, there have been a large number of contributions to AoI analysis.

In package management, AoI in network clouds [42] that transmit out-of-order packets to destinations and with the source node that has the capability to manage [43, 44] the arriving packets was rigorously derived. Ref. [45] controlled the AoI by setting buffer size, deadline, and packet replacement. Three packet management policies were considered with various queues including $M/M/1/2$, $M/M/1/2*$ and $M/M/1/(N+1)*$ [39]. They put forward a lot of effective insights on a lower average age, i.e., the queue can choose to only keep a subset of update packets.

To simplify the analysis, the authors in [43] proposed a new metric, namely peak AoI (PAoI), that characterizes the average maximum elapsed time because the latest received update packet is generated. Focusing on the PAoI, the authors in [39, 46] derived exact PAoI for the $M/G/1$, $M/G/1/1$ and $G/G/1$ models and they showed that PAoI serves as an upper bound for AoI. Age-of-information in the condition of imperfect packet delivery was considered in [47]. Ref. [48] found the optimal scheduling solution for higher information freshness. Ref. [49] showed that the zero-wait policy is not necessarily the optimum policy for freshness or throughput in all real-world scenarios. Currently, a large number of studies are ongoing based on PAoI.

Most of the above studies are based on simple links. There are also some studies that consider AoI for wireless networks. Ref. [50] considered AoI in a general multi-hop network where the update packets do not necessarily arrive to the gateway node in the order of their generation times. Xu et al. [51] calculated and optimized the average PAoI and found the optimal updating frequency in IoT network. The problem of optimizing link scheduling decisions to minimize the expected weighted sum AoI of the clients in the network was researched [52]. In [53], an energy harvesting two-hop network where a source is communicating to a destination through a relay was considered. Based on the Poisson bipolar network, Ref. [54] analyzed the AoI considering spatial distribution, fading, and interference. The author derived average age of information in wireless powered sensor networks in [55].

In sensor networks, energy is a factor limiting access to a smaller average age. Ref. [56] considered the AoI when the time-varying availability of energy at the sender limits the rate of update packet transmissions and showed updates are submitted only when the server is free. The work in [56, 57] showed that a lazy updating policy that introduces inter-update delays is better and Ref. [58] considered threshold-type update policies when the generation of updating opportunities is a Poisson process. AoI in energy harvesting two-hop networks was researched [53]. Ref. [59] generalized the result of [58] to any (integer) battery capacity, and explicitly characterized the threshold structure.

Recently, there are some novel methods applied to the analysis of AoI. Ref. [60] introduced a deep reinforcement learning-based approach that can learn to minimize the AoI with no prior assumptions about network topology. Ref. [61] analyzed the AoI for both the standard ARQ and hybrid ARQ (HARQ) protocols. SHS tools were introduced to extend AoI results to preemptive queues with multiple sources in [62]. Ref. [63] defined soft updates to determine the optimum updating schemes.

Not just in theory, AoI and PAoI are also used in many scenarios, including the vehicular network [64], correlated updates from multiple cameras [65], the context of industrial cyber-physical system [66], CSMA/CA based wireless networks [67], real-life connections [68], TCP/IP links for WiFi, 4G-LTE, 3G, 2G and Ethernet services [69], exchanging systems [70, 71], shared-access channel, differential encoding of temporally correlated updates [72], and game-theoretic approaches to network resource allocation for updating sources [73].

With the development of computers and 5G, the requirements of information freshness are becoming more and more stringent. As a measure of real-time performance, AoI will also receive more attention.

## 5 Challenges

Although many studies dedicated on URLLC including physical layer design (e.g., channel coding, frame structures, HARQ), advanced transmission technology (e.g., grant-free transmission, D2D communications, massive MIMO) as well as network layer optimization have been proposed to meet the stringent requirements in URLLC, there remain many technical challenges to be addressed.

### 5.1 Fundamental constraints

The primary goal of URLLC is to deliver small data payloads (32 to 200 bytes) within a latency of 1 ms. While the design of short packet can essentially decrease the latency, it will cause a severe degradation in channel coding gain. To decrease the control overhead, the control packets should also be short. However, for different payloads, it is unclear the control overhead should be fixed or also flexible. On the other hand, a lower control overhead may not be able to provide a same reliability.

To ensure the 1 ms latency, every component in the transmission should be improved or re-designed, such as the resource grant, TTI duration, and the ACK/NACK mechanism. While ensuring URLLC at a link level in achievable, it is difficult to meet the hard latency constraints at a network level, especially in remote scenarios. This is because network latency may come from intermediate nodes, backhaul links and the core network, the environments of which cannot be easily controlled. Moreover, it is also challenging to investigate the tradeoffs between delay, throughput and reliability in wireless networks including both coding and queuing delays.

Another metric of URLLC, i.e., target BLER, is at least $10^{-5}$ within 1 ms, which is much higher than the current standard. To ensure reliability, more resources are required, e.g., longer block lengths and retransmissions, which increases latency. Thus, the accuracy of channel estimation should be much improved, and advanced diversity schemes with small control overhead to combat the deep fading effect of wireless channels should be introduced.

### 5.2 Coexistence with other services

The future cellular system should be able to provide all three services efficiently, i.e., URLLC, eMBB and mMTC. Different services require different metrics and also different transmission policies. Owing to the low latency and high reliability requirement, the traffic for URLLC has the utmost priority. It is important not to degrade the performance of other services severely while serving the URLLC traffic. Further, the smooth transitions for different transmission policies or techniques also require a new data frame structure and multiplexing schemes for all services. The widely-used static or semi-static resource allocation between URLLC and other services is not efficient in terms of resource utilization. It is therefore necessary to design dynamic multiplexing schemes, for example, eMBB transmission can be preempted if a URLLC packet arrives in the middle of the frame. A key problem of this setting is the joint scheduling of multiple services over different time-scales. A promising approach to achieve this goal is using the network slicing, which is a virtualization that allows multiple logical slices to run on a shared physical network infrastructure. Each logical slice is isolated from each other and can provide customized network features such as bandwidth, latency, and capacity. In [74, 75], the URLLC is incorporated with multicast eMBB to improve the performance of sliced cloud radio access network (C-RAN).

### 5.3 Incorporating new architectures

From the network perspective, URLLC may further benefit from edge computing and edge caching [76]. Edge computing can help offload intensive computation tasks from end users to BSs, such that the end users can have more computing resources to deal with the transmission process. However, edge computing services are incompatible with the current cellular networks. A direct solution is to modify the existing protocol stack to accommodate edge computing services. As it may require substantial network reconstruction, it is essential to consider how to smoothly merge edge computing into the current protocol stack. On the other hand, by proactively caching contents closer to the end users, edge caching can reduce the end to end delay significantly. Despite the potential benefits of edge caching, there is still a long way to go before these benefits can be realized in practice. This is because the edge caching resources are distributed and also limited, compared to the caching resources in the cloud. But the number of possible contents is unlimited and time-varying. It is therefore important to precisely predict what contents are popular such that the edge cache size can be effectively utilized.

## 6 Conclusion

As one of the three application scenarios of 5G, URLLC has some new features such as high reliability, low delay and high availability. URLLC is widely recognized by the industry to be able to apply in the fields of industrial control, factory automation, smart grid, internet of vehicles communication, remote

surgery and other scenarios. In this paper, we illustrated the challenges to enable the cellular system to seamlessly support the integration of URLLC. Meanwhile, we summarized the potential implementation methods of URLLC from both the aspect of the physical layer and the aspect of network architecture. Much work is still required in this area, both on the theory to understand the fundamental limit of URLLC and on meaningful models to fit practical scenarios.

**References**


1 Campbell K, Cruz L, Flanagan B, et al. The 5G Economy: How 5G Will Contribute to the Global Economy. IHS Markit Report, 2019
2 Parvez I, Rahmati A, Guvenc I, et al. A survey on low latency towards 5G: RAN, core network and caching solutions. IEEE Commun Surv Tut, 2018, 20: 3098–3130
3 Sutton G J, Zeng J, Liu R P, et al. Enabling technologies for ultra-reliable and low latency communications: from PHY and MAC layer perspectives. IEEE Commun Surv Tut, 2019, 21: 2488–2524
4 Feng D, She C, Ying K, et al. Toward ultrareliable low-latency communications: typical scenarios, possible solutions, and open issues. IEEE Veh Technol Mag, 2019, 14: 94–102
5 Popovski P, Nielsen J J, Stefanovic C, et al. Wireless access for ultra-reliable low-latency communication: principles and building blocks. IEEE Netw, 2018, 32: 16–23
6 She C, Dong R, Gu Z, et al. Deep learning for ultra-relible and low-latency communications in 6G network. 2020. ArXiv: 2002.11045
7 Hampel G, Li C, Li J. 5G ultra-reliable low-latency communications in factory automation leveraging licensed and unlicensed bands. IEEE Commun Mag, 2019, 57: 117–123
8 Chen H, Abbas R, Cheng P, et al. Ultra-reliable low latency cellular networks: use cases, challenges and approaches. IEEE Commun Mag, 2018, 56: 119–125
9 Ge X. Ultra-reliable low-latency communications in autonomous vehicular networks. IEEE Trans Veh Technol, 2019, 68: 5005–5016
10 Sukhmani S, Sadeghi M, Erol-Kantarci M, et al. Edge caching and computing in 5G for mobile AR/VR and tactile internet. IEEE Multimedia, 2019, 26: 21–30
11 Zhu L, Feng L, Yang Z, et al. Priority-based uRLLC uplink resource scheduling for smart grid neighborhood area network. In: Proceedings of IEEE International Conference on Energy Internet (ICEI), 2019. 510–515
12 3GPP. Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC). Rep. TR 38.824 V1.1.0, Release 16, 2019
13 Bennis M, Debbah M, Poor H V. Ultrareliable and low-latency wireless communication: tail, risk, and scale. Proc IEEE, 2018, 106: 1834–1853
14 Zhong Y, Ge X, Yang H H, et al. Traffic matching in 5G ultra-dense networks. IEEE Commun Mag, 2018, 56: 100–105
15 She C, Yang C, Quek T Q S. Cross-layer optimization for ultra-reliable and low-latency radio access networks. IEEE Trans Wirel Commun, 2018, 17: 127–141
16 Nasrallah A, Thyagaturu A S, Alharbi Z, et al. Ultra-low latency (ULL) networks: the IEEE TSN and IETF DetNet standards and related 5G ULL research. IEEE Commun Surv Tut, 2019, 21: 88–145
17 Jiang X, Shokri-Ghadikolaei H, Fodor G, et al. Low-latency networking: where latency lurks and how to tame it. Proc IEEE, 2019, 107: 280–306
18 Singh B, Tirkkonen O, Li Z, et al. Contention-based access for ultra-reliable low latency uplink transmissions. IEEE Wirel Commun Lett, 2018, 7: 182–185
19 Mahmood N H, Abreu R, Bohnke R, et al. Uplink grant-free access solutions for URLLC services in 5G new radio. In: Proceedings of the 16th International Symposium on Wireless Communication Systems (ISWCS), 2019. 607–612
20 Feng D, Lu L, Yi Y W, et al. Device-to-device communications underlaying cellular networks. IEEE Trans Commun, 2013, 61: 3541–3551
21 Feng D, Lu L, Yi Y W, et al. Device-to-device communications in cellular networks. IEEE Commun Mag, 2014, 52: 49–55
22 She C, Yang C. Available range of different transmission modes for ultra-reliable and low-latency communications. In: Proceedings of IEEE 85th Vehicular Technology Conference (VTC Spring), 2017. 1–5
23 Liu L, Yu W. A D2D-based protocol for ultra-reliable wireless communications for industrial automation. IEEE Trans Wirel Commun, 2018, 17: 5045–5058
24 Panigrahi S R, Bjorsell N, Bengtsson M. Feasibility of large antenna arrays towards low latency ultra reliable communication. In: Proceedings of IEEE International Conference on Industrial Technology (ICIT), 2017. 1289–1294
25 Vu T K, Liu C F, Bennis M, et al. Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks. IEEE Commun Lett, 2017, 21: 2041–2044
26 Zeng J, Lv T, Liu R P, et al. Linear minimum error probability detection for massive MU-MIMO with imperfect CSI in URLLC. IEEE Trans Veh Technol, 2019, 68: 11384–11388
27 Li J, Han Y. Optimal resource allocation for packet delay minimization in multi-layer UAV networks. IEEE Commun Lett, 2017, 21: 580–583
28 Pan C, Ren H, Deng Y, et al. Joint blocklength and location optimization for URLLC-enabled UAV relay systems. IEEE Commun Lett, 2019, 23: 498–501
29 She C, Liu C, Quek T Q S, et al. UAV-assisted uplink transmission for ultra-reliable and low-latency communications. In: Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops), 2018. 1–6
30 Zhong Y, Quek T Q S, Ge X. Heterogeneous cellular networks with spatio-temporal traffic: delay analysis and scheduling. IEEE J Sel Areas Commun, 2017, 35: 1373–1386

31 Zhong Y, Ge X, Han T, et al. Tradeoff between delay and physical layer security in wireless networks. IEEE J Sel Areas Commun, 2018, 36: 1635–1647

32 She C, Chen Z, Yang C, et al. Improving network availability of ultra-reliable and low-latency communications with multi-connectivity. IEEE Trans Commun, 2018, 66: 5482–5496

33 Suer M T, Thein C, Tchouankem H, et al. Multi-connectivity as an enabler for reliable low latency communications-an overview. IEEE Commun Surv Tut, 2020, 22: 156–169

34 Zhang T K, Xu X G, Zhou L, et al. Cache space efficient caching scheme for content-centric mobile ad hoc networks. IEEE Syst J, 2019, 13: 530–541

35 Yu Q, Maddah-Ali M A, Avestimehr A S. Characterizing the rate-memory tradeoff in cache networks within a factor of 2. IEEE Trans Inform Theor, 2019, 65: 647–663

36 Aggarwal V, Chen Y F R, Lan T, et al. Sprout: a functional caching approach to minimize service latency in erasure-coded storage. IEEE/ACM Trans Netw, 2017, 25: 3683–3694

37 Zhong Y, Wang G, Han T, et al. QoE and cost for wireless networks with mobility under spatio-temporal traffic. IEEE Access, 2019, 7: 47206–47220

38 Kaul S, Yates R, Gruteser M. Real-time status: how often should one update? In: Proceedings of IEEE INFOCOM, Orlando, 2012. 2731–2735

39 Kosta A, Pappas N, Angelakis V. Age of information: a new concept, metric, and tool. 2017. https://ieeexplore.ieee.org/document/8187436

40 Kaul S K, Yates R D, Gruteser M. Status updates through queues. In: Proceedings of the 46th Annual Conference on Information Sciences and Systems (CISS), 2012. 1–6

41 Yates R D, Kaul S. Real-time status updating: multiple sources. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Cambridge, 2012. 2666–2670

42 Kam C, Kompella S, Ephremides A. Age of information under random updates. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Istanbul, 2013. 66–70

43 Costa M, Codreanu M, Ephremides A. Age of information with packet management. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Honolulu, 2014. 1583–1587

44 Costa M, Codreanu M, Ephremides A. On the age of information in status update systems with packet management. IEEE Trans Inf Theory, 2016, 62: 1897–1910

45 Kam C, Kompella S, Nguyen G D, et al. Controlling the age of information: buffer size, deadline, and packet replacement. In: Proceedings of IEEE Military Communications Conference, Baltimore, 2016. 301–306

46 Huang L, Modiano E. Optimizing age-of-information in a multi-class queueing system. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Hong Kong, 2015. 1681–1685

47 Chen K, Huang L. Age-of-information in the presence of error. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Barcelona, 2016. 2579–2583

48 He Q, Yuan D, Ephremides A. On optimal link scheduling with min-max peak age of information in wireless systems. In: Proceedings of IEEE International Conference on Communications (ICC), 2016. 1–7

49 Barakat B, Keates S, Wassell I, et al. Is the zero-wait policy always optimum for information freshness (peak age) or throughput? IEEE Commun Lett, 2019, 23: 987–990

50 Bedewy A M, Sun Y, Shroff N B. Age-optimal information updates in multihop networks. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Aachen, 2017. 576–580

51 Xu C, Yang H H, Wang X, et al. Optimizing information freshness in computing-enabled IoT networks. IEEE Internet Things J, 2020, 7: 971–985

52 Kadota I, Uysal-Biyikoglu E, Singh R, et al. Minimizing the age of information in broadcast wireless networks. In: Proceedings of the 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2016. 844–851

53 Arafa A, Ulukus S. Age-minimal transmission in energy harvesting two-hop networks. In: Proceedings of IEEE Global Communications Conference (Globecom), Singapore, 2017. 1–6

54 Hu Y, Zhong Y, Zhang W. Age of information in Poisson networks. In: Proceedings of the 10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, 2018. 1–6

55 Krikidis I. Average age of information in wireless powered sensor networks. IEEE Wirel Commun Lett, 2019, 8: 628–631

56 Bacinoglu B T, Ceran E T, Uysal-Biyikoglu E. Age of information under energy replenishment constraints. In: Proceedings of Information Theory and Applications Workshop (ITA), San Diego, 2015. 25–31

57 Yates R D. Lazy is timely: status updates by an energy harvesting source. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Hong Kong, 2015. 3008–3012

58 Bacinoglu B T, Uysal-Biyikoglu E. Scheduling status updates to minimize age of information with an energy harvesting sensor. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Aachen, 2017. 1122–1126

59 Bacinoglu B T, Sun Y, Uysal-Bivikoglu E, et al. Achieving the age-energy tradeoff with a finite-battery energy harvesting source. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Vail, 2018. 876-880

60 Sert E, Sonmez C, Baghaee S, et al. Optimizing age of information on real-life TCP/IP connections through reinforcement learning. In: Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), 2018. 1–4

61 Ceran E T, Gunduz D, Gyorgy A. Average age of information with hybrid ARQ under a resource constraint. IEEE Trans Wirel Commun, 2019, 18: 1900–1913

62 Yates R D, Kaul S K. The age of information: real-time status updating by multiple sources. IEEE Trans Inform Theor, 2019, 65: 1807–1827

63 Bastopcu M, Ulukus S. Age of information with soft updates. In: Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, 2018. 378–385

64 Kaul S, Gruteser M, Rai V, et al. Minimizing age of information in vehicular networks. In: Proceedings of the 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, Salt Lake City, 2011. 350–358
65 He Q, Dan G, Fodor V. Minimizing age of correlated information for wireless camera networks. In: Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Honolulu,2018. 547–552
66 Sinha D, Roy R. Scheduling status update for optimizing age of information in the context of industrial cyber-physical system. IEEE Access, 2019, 7: 95677–95695
67 Wang M, Dong Y. Broadcast age of information in CSMA/CA based wireless networks. In: Proceedings of 2019 15th International Wireless Communications Mobile Computing Conference (IWCMC), Tangier, 2019. 1102–1107
68 Beytur H B, Baghaee S, Uysal E. Measuring age of information on real-life connections. In: Proceedings of 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, 2019. 1–4
69 Sonmez C, Baghaee S, Ergisi A, et al. Age-of-information in practice: status age measured over TCP/IP connections through WiFi, Ethernet and LTE. In: Proceedings of IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Batumi, 2018. 1–5
70 Hu C, Dong Y. Age of information of two-way data exchanging systems with power-splitting. J Commun Netw, 2019, 21: 295–306
71 Moltafet M, Leinonen M, Codreanu M. Worst case analysis of age of information in a shared-access channel. In: Proceedings of the 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, 2019. 613–617
72 Bhambay S, Poojary S, Parag P. Differential encoding for real-time status updates. In: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, 2017. 1–6
73 Nguyen G D, Kompella S, Kam C, et al. Impact of hostile interference on information freshness: a game approach. In: Proceedings of the 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Paris, 2017. 1–7
74 Tang J, Shim B, Quek T Q S. Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB. IEEE J Sel Areas Commun, 2019, 37: 881–895
75 Anand A, Veciana G, Shakkottai S. Joint scheduling of URLLC and eMBB traffic in 5G wireless networks. In: Proceedings IEEE INFOCOM, Honolulu, 2018. 1970–1978
76 Cao B, Zhang L, Li Y, et al. Intelligent offloading in multi-access edge computing: a state-of-the-art review and framework. IEEE Commun Mag, 2019, 57: 56–62