Daquan Feng, Changyang She, Kai Ying, Lifeng Lai, Zhanwei Hou,
Tony Q.S. Quek, Yonghui Li, and Branka Vucetic

# TOWARD ULTRARELIABLE LOW-LATENCY COMMUNICATIONS

## *Typical Scenarios, Possible Solutions, and Open Issues*

Ultrareliable low-latency communications (URLLC) is one of three emerging application scenarios in 5G new radio (NR) for which physical layer design aspects have been specified. With 5G NR, we can guarantee reliability and latency in radio access networks. However, for communication scenarios where the transmission involves both radio access and wide-area core networks, the delay in radio access networks contributes to only a portion of the end-to-end (E2E) delay. In this article, we outline the delay components and packet loss probabilities in typical URLLC scenarios and formulate the constraints on E2E delay and overall packet loss probability. Then, we summarize possible solutions in the physical, link, and network layers as well as the cross-layer design. Finally, we discuss open issues in prediction and communication codesign for URLLC in wide-area, large-scale networks.

### Achieving URLLC

5G NR considers three new application scenarios, i.e., enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and URLLC [1]. URLLC is crucial for enabling mission-critical services such as factory automation, automated vehicles, remote control, and virtual reality/augmented reality (VR/AR).

There are many open technical hurdles for achieving URLLC; thus, it has attracted significant attention from both the academic and industrial communities. In current long-term evolution (LTE) systems, the transmission time interval (TTI) is 1 ms, which cannot satisfy the E2E delay requirement for URLLC. To reduce its latency, a short frame structure with short channel codes should be considered. However, with short codes, it is very difficult to achieve the ultrahigh reliability requirement. Analyzing and optimizing the transmission delay and decoding error probability in the short block length regime are also very challenging [2].

Aside from transmission delay and decoding error probability, other delay components and delay-bound violation probabilities for scheduling procedure and queueing systems also have significant impacts on E2E performance. For example, in LTE systems, the control signaling for uplink (UL) scheduling leads to a high latency much longer than 1 ms in the control plane [3]. Thus, determining how best to design grant-free access techniques for URLLC deserves further study. Additionally, by using a first-come, first-served (FCFS) scheduling policy, the short packets of URLLC services may need to wait for long packets of eMBB services to be processed. Thus, FCFS may not be the optimal policy for short packets in URLLC services, and other policies should be considered to minimize the E2E delay.

*Digital Object Identifier 10.1109/MVT.2019.2903657*
*Date of publication: 5 April 2019*

1556-6072/19©2019IEEE

IMAGE LICENSED BY INGRAM PUBLISHING

Current techniques in 5G NR [4] focus mainly on achieving the target E2E performance in local area communications where all user equipment (UE) lies in one or several adjacent cells. For many communication scenarios, the network architectures are different. In factory automation, the communication area is limited in a smart factory, while for remote control, the controller and slave can be located on different continents. As a result, the latency in radio access networks contributes only a small portion of the E2E delay; other delay components, such as core network delay over a long-distance and large-scale network and processing delay in the computing systems, may be the dominant components [5]. Therefore, determining how best to improve E2E performance with different network architectures is still a challenging issue.

In this article, we focus on how to guarantee the E2E delay and overall packet loss probability in different communication scenarios, including local area communications, mobile edge computing (MEC) systems, and long-distance large-scale networks.

## E2E Delay and Overall Packet Loss Probability

The factors and delay components that lead to packet loss depend on network architectures. In this section, we first discuss these components and factors in three typical communication scenarios, as illustrated in Figure 1: local area communications, MEC, and wide-area large-scale communications. Then, we provide a general process for formulating quality-of-service (QoS) constraints of URLLC.

### *Local Area Communications*

In local area communications, all UEs are served by a few adjacent access points (APs) interconnected by single-hop fiber backhaul. One typical scenario requiring only local area communications relates to vehicle safety applications, where safety messages are shared among vehicles in close proximity to one another. In this scenario, E2E delay includes UL and downlink (DL) transmission delays (i.e., $D_{t}$,), a queueing delay in the buffer of the APs (i.e., $D_{q}$), and UL access delay (i.e., $D_{a}$); the propagation delay and backhaul delay are negligible because they are much smaller than 1 ms.

To achieve low latency, the transmission delay should be short and so requires short block length channel codes. To achieve ultrahigh reliability, the decoding error probability in the short block length regime, $\varepsilon_{t}$, cannot be ignored [2]. Furthermore, a packet will become useless if the queueing delay or access delay violates the corresponding delay bounds. Therefore, the queueing delay violation probability, $\varepsilon_{q}$, and access delay violation probability, $\varepsilon_{a}$, should be considered for URLLC services.

### *Mobile Edge Computing*

In smart factories or VR/AR applications, UEs may not have sufficient processing capability. In this case, to reduce processing delay at the local computing system, UEs can offload tasks to MEC systems. In MEC systems, all delay components and packet loss factors in local area communications should be considered. Additionally, the processing delay, $D_{p}$, could be comparable to other delay components. If packets are sent to central servers via a multihop backhaul, then the backhaul delay, $D_{b}$, may be dominant. Similar to queueing and access procedures, if packet processing is not finished in time or if the backhaul delay violates the required delay bound, the packet is lost. Thus, both the processing delay violation probability, $\varepsilon_{\mathrm{p}}$, and the backhaul delay violation probability, $\varepsilon_{\mathrm{b}}$, should be taken into account with MEC systems.

### Wide-Area, Large-Scale Networks

Unlike the traditional Internet, which supports real-time audio and video communications, some remote control applications aim to deliver real-time control and tactile feedback (e.g., industrial control, remote driving, or telerobotic surgery). As stated in [5], the long-term goal of a tactile Internet is to enable the sharing of skills globally. In wide-area core networks, additional delays are incurred because of the intermediate data center/cloud. In this case, the overall latency is dictated not only by the radio access networks but also by the backhauls, wireless core networks, and processing in data centers. For example, if the distance between the controller and the slave is 3,000 km, the propagation delay, $D_{\mathrm{g}}$, is roughly 10 ms. To handle this issue, one promising solution is to deploy intelligent MEC systems to predict the mobilities of the controller and slave and transmit their control and feedback information in advance [6].

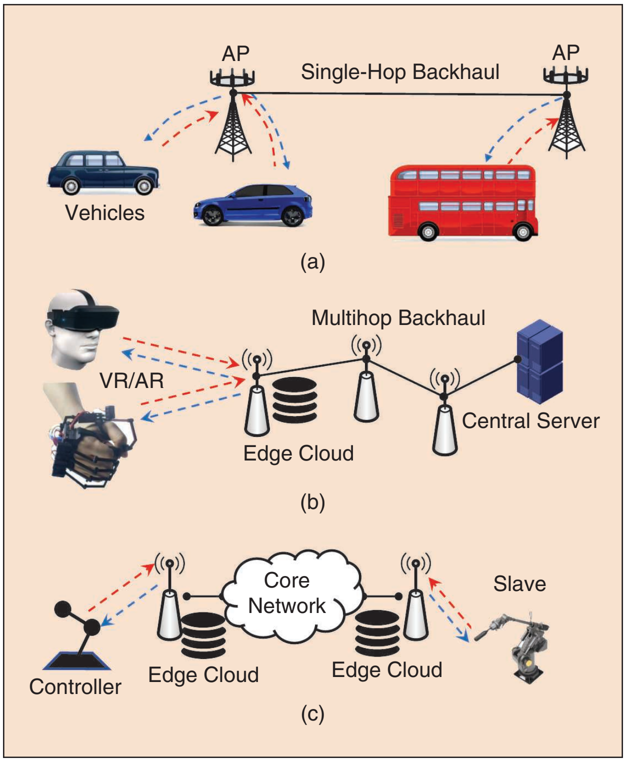


*FIGURE 1* The three typical communication scenarios for URLLC services. (a) Local area communications, (b) MEC, and (c) wide-area, large-scale networks. AP: access point.

### Constraints on E2E Delay and Overall Packet Loss Probability

Denote the requirement of the E2E delay and overall packet loss probability as $D_{\max}$ and $\varepsilon_{\max}$, respectively. Then, the delay and reliability can be satisfied under the following two constraints:

$$D_{\mathrm{t}} + D_{\mathrm{q}} + D_{\mathrm{a}} + D_{\mathrm{p}} + D_{\mathrm{b}} + D_{\mathrm{g}} \leq D_{\max}, \tag{1}$$

$$(1-\varepsilon_{\mathrm{t}})(1-\varepsilon_{\mathrm{q}})(1-\varepsilon_{\mathrm{a}})(1-\varepsilon_{\mathrm{p}})(1-\varepsilon_{\mathrm{b}}) \leq 1-\varepsilon_{\max}. \tag{2}$$

In the following sections, we discuss recent advances in the physical and link layers, network architecture, and cross-layer design, which ensure that the requirements in (1) and (2) are met. Typical applications, possible solutions, and open issues are summarized in Table 1.

## Physical Layer Technologies

Physical layer design is among the most challenging and important issues for the three communication scenarios of URLLC applications. For URLLC in 5G NR, the target user plane latency is 0.5 ms for both UL and DL, while the target reliability is 99.999% success probability for transmitting a packet of 32 bytes within 1 ms [1].

### Flexible Numerology and Frame Structure

In current fourth-generation networks, the TTI is 1 ms; therefore, the TTI should be shortened to meet the URLLC latency requirement. From a frame structure point of view, there are two ways to shorten the TTI. The first is to increase the subcarrier spacing (SCS) so that the symbol duration can be decreased. In 5G NR, the SCS is flexible and is given by $\triangle f = 2^{\mu} \cdot 15\,\mathrm{kHz}$ and

*TABLE 1* ***Applications, possible solutions, and open issues in different communication scenarios.***

| ***Communication Scenarios*** | ***Applications*** | ***Possible Solutions*** | ***Open Issues*** |
|---|---|---|---|
| Local area communications | Road safety applications and autonomous vehicles [1] | 5G NR, grant-free access, and multiconnectivity | Analyzing overhead for channel estimation and correlation of shadowing |
| Edge computing systems | VR/AR and factory automation [6] | Improving scheduling schemes in communication and computing systems | Optimizing communication and computing systems and characterizing E2E delay and reliability |
| Wide-area, large-scale networks | Health care, remote control, and smart grid [5] | Prediction and communication codesign | Designing accurate prediction methods and jointly optimizing prediction and communication systems |

$\mu = 0, 1, 2, 3,$ and 4. Thus, the TTI can be reduced by selecting larger SCS. For example, when the number of orthogonal frequency-division multiplexing (OFDM) symbols in a slot is fixed as 14, the slot duration with 30-kHz SCS is 0.5 ms, while the slot duration with 60-kHz SCS is 0.25 ms. The other way to shorten the TTI is to reduce the number of OFDM symbols in a TTI, i.e., the motivation for introducing a minislot in 5G NR. The number of OFDM symbols in a minislot can be {2, 4, and 7}; as a result, the TTI can be further shortened. For example, the duration of a two-symbol minislot with a 30-kHz SCS is 71.4 $\mu$s, which is much shorter than that of the TTI in LTE.

### *Self-Contained Slot Structure*

In the time-division duplex (TDD) mode, when an AP receives a scheduling request from a UE, it must wait until the next available DL slot to send out a UL grant. However, in a UL-heavy configuration, there are fewer DL slots, and the waiting time can be very long. Similarly, in a DL-heavy configuration, a quick acknowledgment to a DL data reception may not be available.

Thus, in 5G NR, a self-contained slot structure is introduced, where OFDM symbols in a slot can be classified as DL, flexible, or UL. In other words, both directions can be supported within a single slot. In this case, with the help of the self-contained slot structure, the waiting time in TDD systems can be shortened considerably. For example, after receiving DL data at the beginning of a slot, UE can feed back the corresponding acknowledgement at the end of the same slot.

### *Channel Quality Indicator and Modulation and Coding Scheme Table for URLLC*

To guarantee the reliability of data transmission, an appropriate modulation and coding scheme based on the channel quality indicator should be selected from a lookup table to meet the block error rate (BLER) target, i.e., the decoding error probability in (2). The BLER target of eMBB is set as $10^{-1}$, which is the same as for LTE. For URLLC, the target BLER is as low as $10^{-6}$ [1].

On the other hand, to achieve a successful data transmission, the control message must be reliable, whether it is for resource assignment or feedback. Intuitively, there are two basic methods to enhance control reliability: one is to enlarge the control resource; the other is to shorten the size of control information. Both methods help encode the control message with a low coding rate so that reliability is enhanced.

### *Slot Aggregation and Repetition*

To improve reliability, we introduce slot aggregation (for grant-based transmission) and repetition (for grant-free transmission) in NR. The basic idea of slot aggregation and repetition is that an initial transmission of a packet can be followed by automatic repetitions of the same packet in consecutive slots. The aggregation factor (i.e., the number of repetitions) $K$ is configured by the higher layer. $K = 1$ means that there is no aggregation (or repetition) after the initial transmission. According to current NR specification, the largest value of $K$ is 8, which is large enough to guarantee ultrareliability of the data transmission. On the other hand, from a latency perspective, the retransmission timeline is reduced by using repetitions. Retransmission is always grant based, which is time-consuming. However, repetitions are automatically transmitted in the consecutive slots without waiting for any grant or retransmission feedback, thereby helping to reduce latency.

**A PACKET WILL BECOME USELESS IF THE QUEUEING DELAY OR ACCESS DELAY VIOLATES THE CORRESPONDING DELAY BOUNDS.**

## Link Layer Design

In this section, we focus on link layer design and consider the scenarios that include random packet arrival processes. To reduce latency, grant-free access for UL transmission and DL scheduling policies in communication and processing systems are summarized. Additionally, to improve reliability, device-to-device (D2D) communications, relay systems, cellular links, and multiconnectivity are discussed.

### *Grant-Free Access for UL Transmission*

With the current LTE protocol, when a UE has a packet to transmit, it first uploads a scheduling request to the AP. Then, the AP sends a transmission grant to the UE. Finally, the UE can upload its packet. Such a UL scheduling procedure leads to a lengthy access delay, which can be mitigated by grant-free access.

Reserving dedicated bandwidth for each UE is a natural way to avoid access delay; however, such a method is only suitable for UEs with a high packet arrival rate. Reserving bandwidth for each UE with low packet arrival rate leads to very low bandwidth usage efficiency. To address this issue, a contention-based access procedure has been studied in [7], i.e., the slotted ALOHA access scheme. With contention-based access, the total bandwidth is divided into multiple channels. In each slot, the UEs that need to send packets choose one of the channels randomly for data transmission. If more than one UE chooses the same channel, then the transmissions will fail.

According to the experiment in [3], the arrival processes in some applications are very bursty. As illustrated in Figure 2, the arrival rate of a bursty arrival process switches between a high-traffic and low-traffic state. To avoid high-collision probability and save bandwidth, the

**AFTER RECEIVING DL DATA AT THE BEGINNING OF A SLOT, UE CAN FEED BACK THE CORRESPONDING ACKNOWLEDGEMENT AT THE END OF THE SAME SLOT.**

authors in [8] first classify the arrival processes into high- and low-traffic states. Then, dedicated bandwidth is reserved for UEs in the high-traffic state, and the slotted ALOHA access scheme is applied for UEs in the low-traffic state. To guarantee the reliability requirement, classification errors should be considered [8].

### DL Scheduling Policies in Communication and Computing Systems

In traditional communication systems, the packets to different destinations are waiting in different queues at the buffer of the AP, e.g., the individual FCFS server shown in Figure 3(a). Such a policy can guarantee the QoS of each user; however, the resource utilization efficiency is low because the resources allocated to different users cannot be shared even though some users' queues are empty. To improve resource utilization efficiency, the statistical multiplexing server in Figure 3(b) can be used; here, packets of different users stay in one queue. As proved in [9], if the arrival of each user follows a Poisson process and the packet sizes are identical (e.g., each short packet in URLLC contains 20 bytes [1]), then the statistical multiplexing server can guarantee the QoS of all the packets with less total bandwidth compared to adopting an individual server for each user.

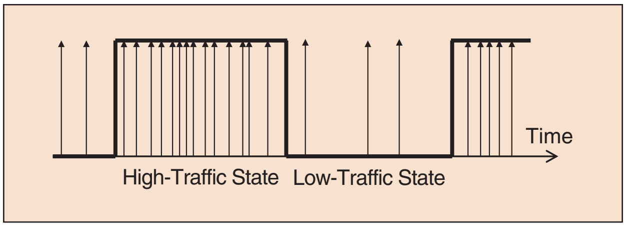


FIGURE 2 The bursty packet arrival process.

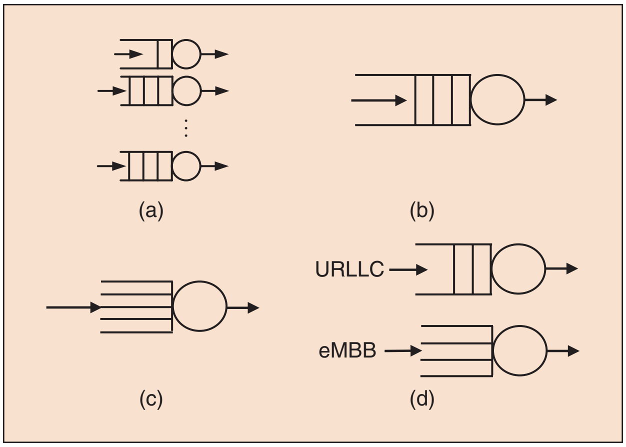


FIGURE 3 Typical scheduling policies: the (a) individual FCFS server, (b) statistical multiplexing FCFS server, (c) PS server, and (d) service-aware server.

In practical systems, the transmission/processing time of different kinds of packets can be very diverse. Because the packet size in eMBB services is much larger than in URLLC services, the transmission/processing time of long packets in eMBB services is much longer than that of short packets in URLLC services. As a result, if we use FCFS servers, the short packets that arrive at the server following a long packet must wait for a long period of time. To avoid this situation, one solution proposes the processor-sharing (PS) server, as illustrated in Figure 3(c). In the PS server, the total service ability is equally allocated to all packets in the buffer [10]. In this way, the short packets do need to not wait for long packets to be processed. Furthermore, if the server is aware of the packets in different services, then it is possible to design different scheduling policies for different types of services [e.g., the service-aware server shown in Figure 3(d)].

### D2D, Relay, and Cellular Links for URLLC

In some applications of URLLC, such as factory automation and vehicle networks, each device transmits short packets to nearby devices. For these short-distance communication scenarios, D2D communications may outperform cellular links. However, when using D2D communication in a URLLC scenario, interference should be avoided. One possible solution is using APs to manage radio resources and sending data packets via D2D links.

Considering that D2D communications are limited in range, relay systems are applied in [11]. The results in [11] demonstrate that these relay systems can achieve higher throughput or lower queueing delay compared to that of direct transmission in both noise-limited and interference-limited scenarios.

To further improve reliability, one important technique is multiconnectivity [12], which involves transmitting one packet over all of the multiple parallel links, such as D2D links, relay, and cellular links. The results in [12] show that the achieved reliability decreases with the cross-correlation of shadowing among parallel links. However, how to analyze the impact of cross-correlation of shadowing on reliability remains an open problem.

## Novel Network Layer Design

Existing cellular network architecture is mainly designed to meet the requirements of conventional mobile broadband services. However, this architecture cannot support diversified 5G services. Thus, novel mobile computing frameworks and network architecture techniques, such as network slicing, software-defined networking (SDN), network function visualization (NFV), and self-organizing networks (SON) are attracting considerable attention. In this section, we introduce recent advances of these techniques that are used for URLLC services.

### Network Slicing with SDN/NFV

Network slicing is a fundamental technology for future networks and provides diversified services simultaneously over the same physical infrastructure [13]. It allows the network to build multiple logical subnetworks with reserved resources for different application scenarios, e.g., eMBB, mMTC, and URLLC. In this context, URLLC service can avoid the interruptions caused by other services that share the same resources and thus help enhance QoS and user experience. However, because of traffic fluctuations and channel variability, strict isolation and sharing among multiple slices faces a significant challenge and may lead to low resource utilization efficiency.

SDN and NFV are two pillars that support multivirtual network slicing in 5G. In particular, SDN separates the control plane from the forwarding plane and offers centralized network flow management that simplifies scheduling and resource allocation. On the other hand, NFV decouples network functions from dedicated hardware to provide programmability and flexibility over the entire network. With the integration of SDN and NFV, network operators can provide efficient, scalable, and flexible network slicing service configurations on demand. Therefore, network slicing with SDN and NFV can significantly improve overall network performance including delay and reliability in radio access networks, backhauls, and core networks [14]. The authors of [15] show that the SDN-based network architecture can achieve an up to 75% performance improvement in E2E latency. In [16], the authors propose a two-level medium-access-control scheduling framework for a slicing-enabled 5G network. Using dynamic slice management, the stringent requirements for URLLC can be guaranteed.

### Reducing Latency With MEC

MEC is another promising solution that reduces latency when processing tasks for URLLC services. In [17], MEC is considered for integration with SDN and NFV to deal with the service disruption incurred by user mobility. The authors demonstrate that distributed and virtualized network provisioning can effectively reduce latency and improve resiliency. In [18], the tradeoff between power and delay in MEC is studied, with computation and transmit power minimized by optimizing task offloading and resource allocation. Because of the FCFS scheduling policy discussed previously, the latency of short packets is not optimized. Determining how best to optimize task offloading and scheduling policy subject to the QoS requirements of URLLC, eMBB, and mMTC deserves further study.

**NETWORK SLICING IS A FUNDAMENTAL TECHNOLOGY FOR FUTURE NETWORKS AND PROVIDES DIVERSIFIED SERVICES SIMULTANEOUSLY OVER THE SAME PHYSICAL INFRASTRUCTURE.**

### Self-Organizing Networks

Networking slicing, SDN, and NFV can improve network scalability and flexibility. However, they also result in much more complicated network management and configuration, which may depress the QoS and user experience. Thus, it is critical to adopt SON management mechanisms, which provide intelligence, automation, and distributed management and optimization [19]. By taking advantage of the rapid progress of big data processing and machine-learning technologies, the European Union's 5G Public Private Partnership project proposes a catalog-driven network management system that enables the smart deployment of service. Nevertheless, how best to guarantee the QoS requirement of URLLC in SON deserves further study.

## Improving E2E Performance With Cross-Layer Design

Considering that each layer of the protocol stack has an inherent interdependence on other layers, cross-layer resource management has the potential to improve E2E delay and overall reliability. For example, transmission, queueing, and routing delays all depend on the physical, link, and network layers, respectively. By optimizing the delay components subject to the E2E delay constraint in (1), we can achieve better resource utilization efficiency. In this section, we describe how best to save bandwidth or transmit power by using a cross-layer design.

### A Different Conclusion Obtained From Cross-Layer Design

In the physical layer design, it is well known that there are tradeoffs among physical layer resources, e.g., transmission time, bandwidth, and transmit power. As shown in Figure 4(a), if we double the transmission duration, then only half the bandwidth is required if the rate of the channel code remains constant. Furthermore, the transmit power used to achieve the same SNR is also halved.

However, the conclusion is different from a cross-layer perspective. We consider an FCFS queueing system and assume that, to guarantee a queueing delay bound, $D_{\mathrm{q}}$, and queueing delay violation probability, $\varepsilon_{\mathrm{q}}$, the required service rate, referred to as e*ffective bandwidth* [9], is $E_{\mathrm{B}} = 2$ (packets/frame). If the transmission duration of each packet is 1 frame, then two packets are transmitted simultaneously. As shown in Figure 4(b), if the transmission duration of each packet is 2 frames, then, to achieve the required service rate, i.e., 2 (packets/frame), the number of packets transmitted simultaneously is four. As a result, although the bandwidth for each packet is halved, the total bandwidth and total transmit power remain unchanged. Therefore, increasing the transmission duration does not help reduce bandwidth or transmit power but does lead to extra transmission delay. Consequently, the optimal transmission duration that minimizes the required bandwidth subject to the constraints on transmission and queueing delays and overall packet loss probability is 1 frame [9].

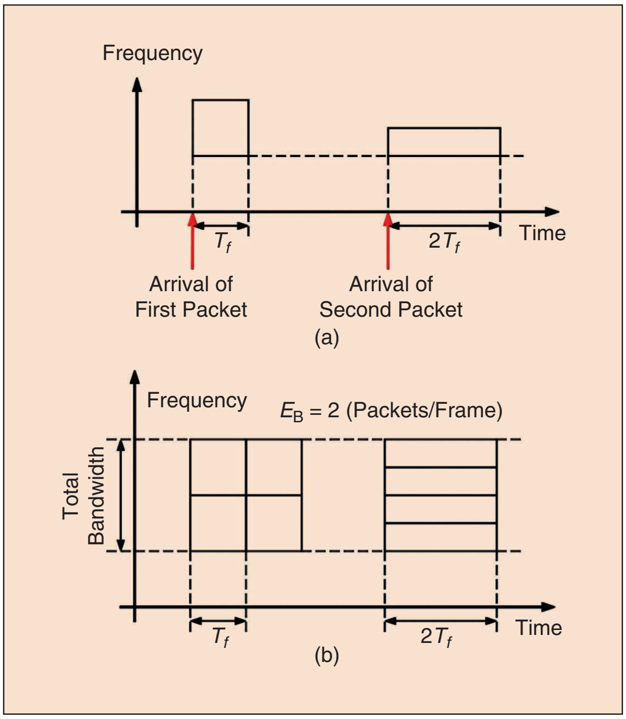


FIGURE 4 An illustration showing the importance of the cross-layer design. (a) The physical layer design and (b) the cross-layer design.

### Useful Insights in Cross-Layer Design

In radio access networks, the most challenging issue for cross-layer design is how to obtain an optimal solution subject to the requirements related to transmission and queueing delays as well as different packet loss probabilities. In [20], the required transmit power is minimized by jointly optimizing the probability of decoding errors, queueing delay violations, and packet dropping over deep-fading wireless channels, subject to the overall packet loss probability requirement. The results in Figure 5 indicate that only a ~2–5% power saving can be obtained by optimizing packet loss probabilities when the number of antennas at the AP is more than eight. A near-optimal solution is to set all packet loss components in (2) as equal. Furthermore, the UL and DL transmission delays as well as the queueing delay are optimized, subject to the E2E delay requirement in [9]. The results in Figure 6 show that, by optimizing these three delay components, approximately half the bandwidth can be saved. This is because the required resources are very sensitive to the delay components but less sensitive to the packet loss probabilities.

## Toward Wide-Area, Large-Scale Networks: Prediction and Communication Codesign

The propagation delay, $D_g$, will be higher than 1 ms as long as the communication distance is longer than 300 km. It is therefore impossible to achieve a 1-ms E2E delay with only physical layer technologies. Inspired by existing studies on mobility prediction [21]–[23], we propose a prediction and

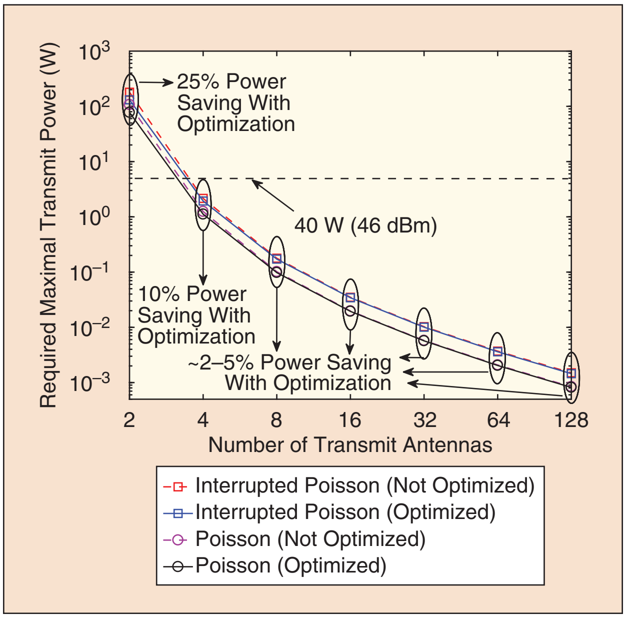


FIGURE 5 The performance gain achieved from optimizing the packet loss probabilities [20].

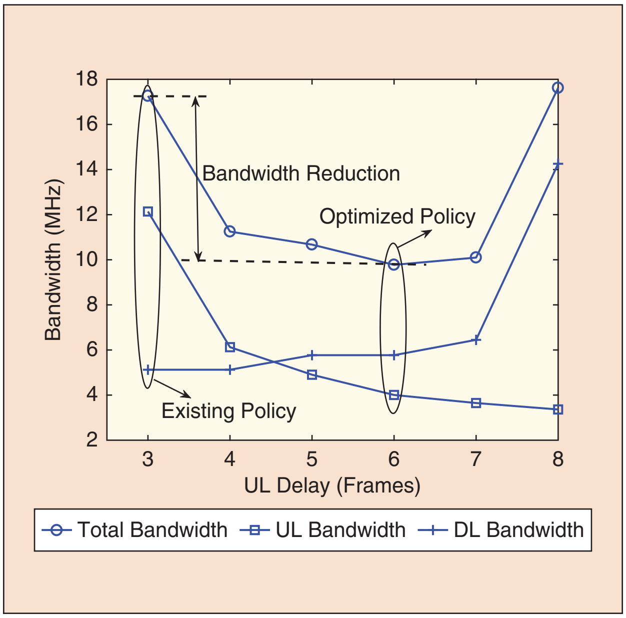


FIGURE 6 The performance gain achieved from optimizing the delay components [9].

communication codesign method to handle this issue. The idea is to predict the device's movement and send the predicted information in advance. Assuming that the system can predict the mobilities of the controller and slave (i.e., $T_e$ s) in advance, the delay in the core network experienced by the controller and slave can be reduced. In this case, prediction errors will lead to packet loss, and we denote $\varepsilon_c$ as the packet loss probability due to prediction errors. The UL transmission with prediction is shown in Figure 7. At time $t$, the device predicts its future location and sends the predicted information to the remote controller. With communication delay, the E2E delay experienced by the remote controller is the same as what is shown on the right-hand side of (1). If the sum of the delay components in the communication system is equal to the prediction time, then it is possible to achieve zero-latency.

However, there are three open issues. First, intuitively, there is a tradeoff between the prediction time, $T_e$, and the prediction error probability, i.e., $\varepsilon_c = \Pr\{|S(t+T_e) - \hat{S}(t+T_e)|\}$. Deciding how best to design an accurate prediction algorithm that achieves low prediction error probability with long prediction time deserves further study. Possible solutions include using model-based methods with Markov chain or first-order autoregressive models as well as data-driven methods such as linear regression and neural networks. Second, latency in the communication system is a random variable that depends on wireless channel fading, queueing, routing, and network congestion. Determining how best to satisfy the constraint on the probability of the experienced delay violating the delay bound is an open problem. Third, how best to optimize prediction time to minimize the overall packet loss probability for a given prediction algorithm and a communication system remains unclear. To this end, a prediction and communication codesign is necessary.

**DECIDING HOW BEST TO DESIGN AN ACCURATE PREDICTION ALGORITHM THAT ACHIEVES LOW PREDICTION ERROR PROBABILITY WITH LONG PREDICTION TIME DESERVES FURTHER STUDY.**

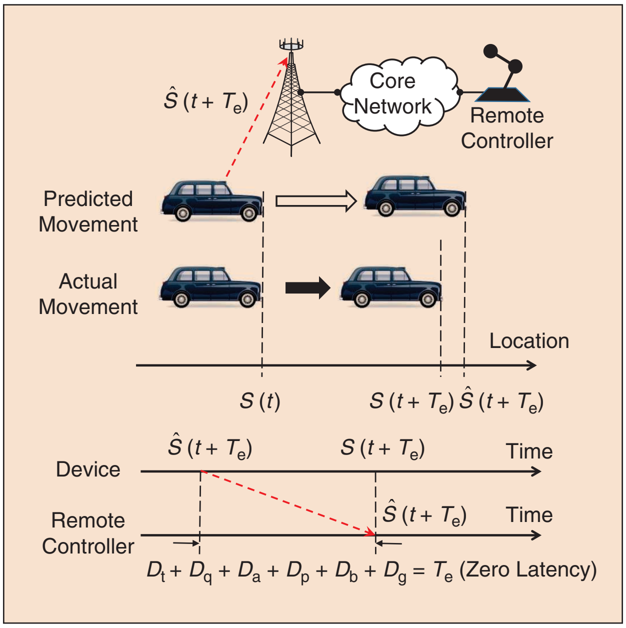


FIGURE 7 An illustration of prediction and communication.

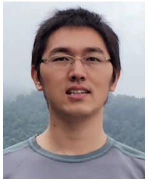
## Conclusions

In this article, we discussed the delay components and packet loss probabilities in three typical communication scenarios for URLLC. Then, we summarized the possible solutions and techniques in the physical and link layers as well as the network architecture design aspects for URLLC. The solutions from each of these three layers are important for enabling URLLC. However, without cross-layer optimization, the separate optimization in the three aspects cannot obtain the global optimal solution and may lead to incorrect conclusions. Motivated by this fact, we presented some optimization results in cross-layer resource management. Finally, we outlined the basic idea of prediction and communication codesign for wide-area, large-scale networks and discussed some open issues.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant 61701317, the Young Elite Scientists Sponsorship Program by China Association for Science and Technology under grant 2018QNRC001, the Guangdong Natural Science Foundation under grant 2017A030310371, the Shenzhen Basic Research Program under grant JCYJ20170302150006125, the Start-Up Fund of Shenzhen University under grant 2017076, the "Tencent Rhinoceros" Birds—Scientific Research Foundation for Young Teachers of Shenzhen University, the Start-Up Fund of Peacock Project, the Singapore University of Technology and Design (SUTD) and Zhejiang University of China (ZJU) Research Collaboration under grant SUTD-ZJU/Research/01/2016 and SUTD-ZJU/RES/05/2016, the Australian Research Council under grant DP150104019 and DP190101988, and the University of Sydney.

## Author Information

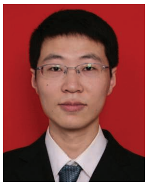


***Daquan Feng*** (fdquan@gmail.com) is an assistant professor with the Guangdong Key Laboratory of Intelligent Information Processing, College of Information Engineering, Shenzhen University, China. His research interests include ultrareliable low-latency communications, long-term evolution in spectrum communications, and massive Internet of Things networks. He is a Member of the IEEE.

***Changyang She*** (changyang.she@sydney.edu.au) is a postdoctoral research associate at the University of Sydney, Australia. His research interests include ultra-reliable and low-latency communications, the tactile Internet, big data for resource allocation in wireless networks, and energy-efficient transmission in 5G communication systems. He is a Member of the IEEE.

***Kai Ying*** (yingkai0301@gmail.com) is a senior researcher at Sharp Laboratories of America, Camas, Washington. His research interests include nonlinear signal processing and ultrareliable and low-latency communications. He is a Member of the IEEE.

***Lifeng Lai*** (llf080915@163.com) is currently pursuing his M.S. degree in electronic and communication engineering at Shenzhen University, China. His research interests include ultrareliable low-latency communications and long-term evolution in unlicensed spectrum communications.



***Zhanwei Hou*** (zhou9027@uni.sydney.edu.au) is a postdoctoral fellow at the University of Sydney, Australia. His research interests involve the tactile Internet, ultra-reliable and low-latency communications, and machine learning.

***Tony Q.S. Quek*** (tonyquek@sutd.edu.sg) is a tenured associate professor with Singapore University of Technology and Design (SUTD). He also serves as the acting head of the Information Systems Technology and Design Pillar and deputy director of the SUTD-Zhejiang University, Innovation, Design and Entrepreneurship Alliance. His research interests include wireless communications and networking, the Internet of Things, network intelligence, wireless security, and big data processing. He is a Fellow of the IEEE.

***Yonghui Li*** (yonghui.li@sydney.edu.au) is a professor in the School of Electrical and Information Engineering, University of Sydney, Australia. His research interests include wireless communications, with a focus on multiple-input, multiple-output; millimeter-wave communications; machine-to-machine communications; coding techniques; and cooperative communications. He is a Fellow of the IEEE.

***Branka Vucetic*** (branka.vucetic@sydney.edu.au) is an Australian Research Council laureate fellow and director of the Centre of Excellence for IoT and Telecommunications at the University of Sydney, Australia. Her research work is in wireless networks, with a focus on communication systems design for millimeter-wave frequency bands, and in the Internet of Things, with a focus on providing wireless connectivity for mission-critical applications. She is a fellow of the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science as well as a Fellow of the IEEE.

## References

[1] 3GPP, "Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)," 3rd Generation Partnership Project, Sophia Antipolis, France, Rep. TR 38.824 V1.1.0, Release 16, 2019.

[2] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[3] M. Condoluci, T. Mahmoodi, E. Steinbach, and M. Dohler, "Soft resource reservation for low-delayed teleoperation over mobile networks," *IEEE Access*, vol. 5, pp. 10,445–10,455, May 2017.

[4] 3GPP, "Study on new radio (NR) access technology; physical layer aspects," 3rd Generation Partnership Project, Sophia Antipolis, France, Rep. TR 38.802, v2.0.0, Release 14, 2017.

[5] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the tactile Internet: Haptic communications over next generation 5G cellular networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 82–89, 2017.

[6] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Towards haptic communications over the 5G tactile Internet," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3034–3059, 2018.

[7] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182–185, Oct. 2017.

[8] Z. Hou, C. She, Y. Li, T. Q. S. Quek, and B. Vucetic, "Burstiness aware bandwidth reservation for ultra-reliable and low-latency communications (URLLC) in tactile Internet," *IEEE J. Select. Areas Commun.*, vol. 36, no. 11, pp. 2401–2410, 2018.

[9] C. She, C. Yang, and T. Q. S. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2266–2280, 2018.

[10] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge, MA: Cambridge Univ. Press, 2013.

[11] Y. Hu, M. C. Gursoy, and A. Schmeink, "Relaying-enabled ultra-reliable low-latency communications in 5G," *IEEE Netw.*, vol. 32, no. 2, pp. 62–68, 2018.

[12] C. She, Z. Chen, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Improving network availability of ultra-reliable and low-latency communications with multi-connectivity," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5482–5496, 2018.

[13] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, 2017.

[14] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: Ran, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 32, no. 4, pp. 3098–3130, 2018.

[15] R. Trivisonno, R. Guerzoni, I. Vaishnavi, and D. Soldani, "Towards zero latency software defined 5G networks," in *Proc. IEEE ICC Workshop*, 2015, pp. 2566–2571.

[16] A. Ksentini, P. A. Frangoudis, A. PC, and N. Nikaein, "Providing low latency guarantees for slicing-ready 5G systems via two-level MAC scheduling," *IEEE Netw.*, vol. 32, no. 6, pp. 116–123, 2018.

[17] R. Ford, A. Sridharan, R. Margolies, R. Jana, and S. Rangan, "Provisioning low latency, resilient mobile edge clouds for 5G," in *Proc. IEEE INFOCOM Workshops*, 2017, pp. 169–174.

[18] C. F. Liu, M. Bennis, and H. V. Poor, "Latency and reliability-aware task offloading and resource allocation for mobile edge computing," in *Proc. IEEE GLOBECOM Workshops*, 2017, pp. 1–7.

[19] 5GPPP Architecture Working Group, "View on 5G architecture," 5GPP Infrastructure Public Private Partnership, White Paper, 2017.

[20] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, 2018.

[21] S. K. Dandapat, S. Pradhan, N. Ganguly, and R. R. Choudhury, "Sprinkler: Distributed content storage for just-in-time streaming," in *Proc. ACM CellNet*, 2013, pp. 19–24.

[22] F. Zhang et al., "EdgeBuffer: Caching and prefetching content at the edge in the MobilityFirst future Internet architecture," in *Proc. IEEE WoWMoM*, 2015, pp. 1–9.

[23] E. Ozfatura and D. Gündüz, "Mobility and popularity-aware coded small-cell caching," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 288–291, 2018.

***VT***