Ultra-Reliable Low Latency Cellular Networks:

Use Cases, Challenges and Approaches

He Chen, Rana Abbas[1], Peng Cheng[1], Mahyar Shirvanimoghaddam[1], Wibowo Hardjawana[1], Wei Bao[1], Yonghui Li, and Branka Vucetic

The University of Sydney, NSW 2006, Australia
Email: firstname.lastname@sydney.edu.au

Abstract-**The fifth-generation cellular mobile networks are expected to support mission critical ultra-reliable low latency communication (uRLLC) services in addition to the enhanced mobile broadband applications. This article first introduces three emerging mission critical applications of uRLLC and identifies their requirements on end-to-end latency and reliability. We then investigate the various sources of end-to-end delay of current wireless networks by taking the 4G Long Term Evolution (LTE) as an example. Subsequently, we propose and evaluate several techniques to reduce the end-to-end latency from the perspectives of error control coding, signal processing, and radio resource management. We also briefly discuss other network design approaches with the potential for further latency reduction.**

I. Introduction

The growth of wireless data traffic over the past three decades has been relentless. The upcoming fifth-generation (5G) of wireless cellular networks is expected to carry 1000 times more traffic [1] while maintaining high reliability. Another critical requirement of 5G is ultra-low latency – the time required for transmitting a message through the network. The current fourth-generation (4G) wireless cellular networks have a nominal latency of about 50ms; however, this is currently unpredictable and can go up to several seconds [2]. Moreover, it is mainly optimized for mobile broadband traffic with target block error rate (BLER) of 10^{-1} before re-transmission.

---

[1] These authors contributed equally to this article.

There is a general consensus that the future of many industrial control, traffic safety, medical, and internet services depends on wireless connectivity with guaranteed consistent latencies of 1ms or less and exceedingly stringent reliability of BLERs as low as 10^{-9} [3]. While the projected enormous capacity growth is achievable through conventional methods of moving to higher parts of the radio spectrum and network densifications, significant reductions in latency, while guaranteeing an ultra-high reliability, will involve a departure from the underlying theoretical principles of wireless communications.

II. Emerging uRLLC Applications

In this section, we introduce three emerging mission-critical applications and identify their latency and reliability requirements.

A. Tele-surgery

The application of uRLLC in tele-surgery has two main use cases [4]: (1) remote surgical consultations, and (2) remote surgery. The remote surgical consultations can occur during complex life-saving procedures after serious accidents with patients having health emergency that cannot wait to be transported to a hospital. In such cases, first-responders at an accident venue may need to connect to surgeons in hospital to get advice and guidance to conduct complex medical operations. On the other hand, in a remote surgery scenario, the entire treatment procedure of patients is executed by a surgeon at a remote site, where hands are replaced by robotic arms. In these two use cases, the communication networks should be able to support the timely and reliable delivery of audio and video streaming. Moreover, the haptic feedback enabled by various sensors located on the surgical equipment is also needed in remote surgery such that the surgeons can feel what the robotic arms are touching for precise decision-

making. Among these three types of traffic, it is haptic feedback that requires the tightest delay requirement with the end-to-end round trip times (RTTs) lower than 1ms [4]. In terms of reliability, rare failures can be tolerated in remote surgical consultations, while the remote surgery demands an extremely reliable system (BLER down to 10^{-9}) since any noticeable error can lead to catastrophic outcomes.

B. *Intelligent Transportation*

The realization of uRLLC can empower several technological transformations in transportation industry [5], including automated driving, road safety and traffic efficiency services, etc. These transformations will get cars fully connected such that they can react to increasingly complex road situations by cooperating with others rather than relying on its local information. These trends will require information to be disseminated among vehicles reliably within extremely short time duration. For example, in fully automated driving with no human intervention, vehicles can benefit by the information received from roadside infrastructure or other vehicles. The typical use cases of this application are automated overtake, cooperative collision avoidance and high density platooning, which require an end-to-end latency of 5–10ms and a BLER down to 10^{-6} [5].

C. *Industry Automation*

uRLLC is one of the enabling technologies in the fourth industrial revolution [6]. In this new industrial vision, industry control is automated by deploying networks in factories. Typical industrial automation use cases requiring uRLLC include factory, process, and power system automation. To enable these applications, an end-to-end latency lower than 0.5ms and an exceedingly high reliability with BLER of 10^{-9} should be supported [3]. Traditionally, industrial control systems are mostly based on wired networks because the existing wireless technologies

cannot meet the industrial latency and reliability requirements. Nevertheless, replacing the currently used wires with radio links can bring substantial benefits: (1) reduced cost of manufacturing, installation and maintenance; (2) higher long-term reliability as wired connections suffer from wear and tear in motion applications; (3) inherent deployment flexibility.

Other possible applications of uRLLC include Tactile Internet, augmented/virtual reality, fault detection, frequency and voltage control in smart grids.
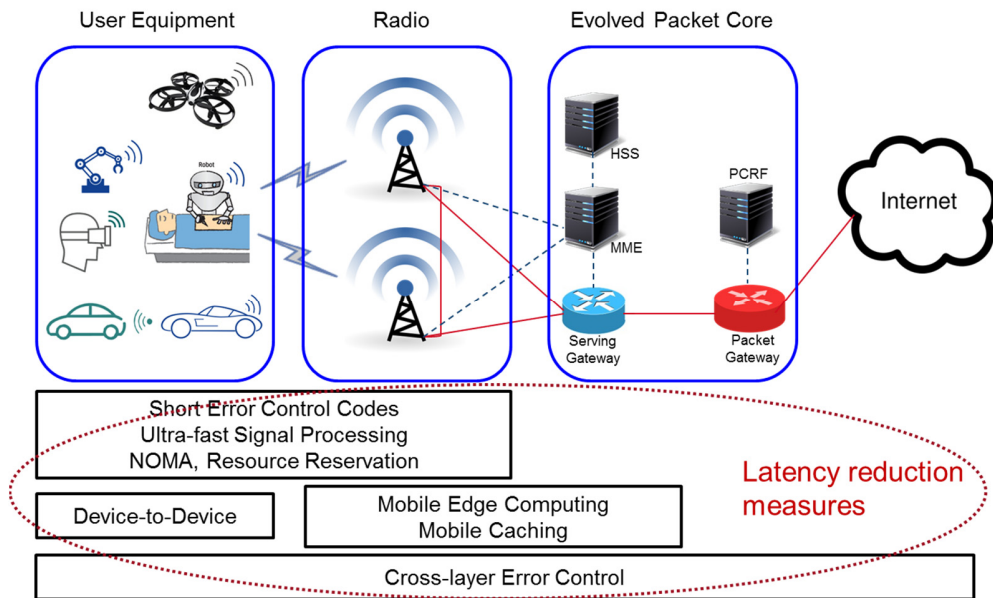

Fig. 1. Architecture of 4G LTE network with representative mission-critical user equipment. The bottom part lists various potential measures towards latency reduction in different parts.

III. Latency Sources in Cellular Networks

Cellular networks are complex systems with multiple layers and protocols, as depicted in Fig. 1. The duration of a data block at the physical layer is a basic delay unit which gets multiplied over higher layers and thus causes a considerable latency in a single link. On the other hand, protocols at higher layers and their interactions are significant sources of delay in the whole network. Latency varies significantly as a function of multiple parameters, including the transmitter–

receiver distance, wireless technology, mobility, network architecture, and the number of active network users.

TABLE I

Various delay sources of an LTE system (Release 8) in the uplink and downlink

Delay Component	Description	Time (ms)
Grant acquisition	A user connected and aligned to a base station will send a Scheduling Request (SR) when it has data to transmit. The SR can only be sent in an SR-valid Physical Uplink Control Channel (PRCCH). This component characterizes the average waiting time for a PRCCH.	5ms
Random Access	This procedure applies to the users not aligned with the base station. To establish a link, the user initiates an uplink grant acquisition process over the random access channel. This process includes preamble transmissions and detection, scheduling, and processing at both the user and the base station.	9.5ms
Transmit time interval	The minimum time to transmit each packet of request, grant or data	1ms
Signal processing	The time used for the processing (e.g., encoding and decoding) data and control	3ms
Packet retransmission in access network	The (uplink) hybrid automatic repeat request process delay for each retransmission	8ms
Core network/Internet	Queueing delay due to congestion, propagation delay, packet retransmission delay caused by upper layer (e.g., TCP)	Vary widely

The latency components of the LTE networks have been systematically evaluated and quantified in [7]. Latencies for various radio access network algorithms and protocols in data transmission from a user to the gateway (i.e., uplink) and back (i.e., downlink) are summarized in Table I. The two most critical sources of delay in radio access networks are the link establishment (i.e., grant acquisition or random access) and packet retransmissions caused by channel errors and

congestion. Another elementary delay component is the transmit time interval (TTI), defined as the minimum data block length, which is involved in each transmission of grant, data, and retransmission due to errors detected in higher layer protocols.

According to Table I, after a user is aligned with the base station, its total average radio access delay for an uplink transmission can be up to 17ms excluding any retransmission. The delay for a downlink transmission is around 7.5ms, which is lower than that of the uplink since no grant acquisition process is needed in the downlink. The overall end-to-end latency in cellular networks is dictated not only by the radio access network but also includes delays of the core network, data center/cloud, Internet server and radio propagation. It increases with the transmitter-receiver distance and the network load. As shown by the experiment conducted in [8], at least 39ms is needed to contact the core network gateway, which connects the LTE system to the Internet, while a minimum of 44ms is required to get response from the Google server. As the number of users in the network rises, the delay goes up, due to more frequent collisions in grant acquisition and retransmissions caused by inter-user interference.

In the subsequent sections, we will consider novel approaches that could be implemented at various cellular network layers (as depicted in the bottom part of Fig. 1) to support ultra-low latency services.

IV. Short Error Control Codes

In traditional communication systems, very long low-density parity check (LDPC) or turbo codes are used to achieve near error-free transmissions, as long as the data rate is below the Shannon channel capacity. Since the network latency is significantly affected by the size of data blocks, short codes are a prerequisite for low delays; but the Shannon theoretical model breaks down for

short codes. A recent Polyansky-Poor-Verdu (PPV) analysis of channel capacity with finite block lengths [9] has provided the tradeoffs between delays, throughput, and reliability on Gaussian channels and fixed rate block codes, by introducing a new fundamental parameter called 'channel dispersion'; this analysis shows that there is a severe capacity loss at short block-lengths.

There are no known codes that achieve the PPV limit. Low-density parity check (LDPC) codes and polar codes have been reported to achieve almost 95% of the PPV bound at block error rates as low as 10^{-7} for block lengths of a few hundred symbols [10]. However, their main drawback is the large decoding latency. On the other hand, convolutional codes provide fast decoding as a block can be decoded as it is being received and can achieve BLERs as low as 10^{-9}. Note that as the signal-to-noise ratio (SNR) in wireless channels varies over time and frequency due to fading, these low BELRs can only be achieved at very high SNRs (as high as 90 dB) over point to point channels. To address this issue, these error control codes need to be augmented by some form of diversity such as implementing multi-antenna techniques.

As long fixed rate codes achieve the Shannon capacity limit for one signal-to-noise ratio (SNR) only, today's wireless networks use adaptive schemes, which select a code from a large number of fixed rate codes, to transmit data at the highest possible rate for a specified reliability and estimated channel state information (CSI). The problem is the inevitable latency increase due to complex encoding and decoding algorithms, the time required to estimate the CSI at the receiver, the feedback of CSI back to the transmitter, code rate and modulation selection process in the transmitter, and block length.

In this context, self-adaptive codes appear as a promising solution to uRLLC. Self-adaptive codes, also known as rateless codes, can adapt the code rate to the channel variations by sending

an exact amount of coded symbols needed for successful decoding. This self-adaptation does not require any channel state information at the transmitter side, thus eliminating the channel estimation overhead and delay. While there are some research results on rateless codes for the short block length regime, they are all on binary codes, and their extension to the real domain is not straight-forward.
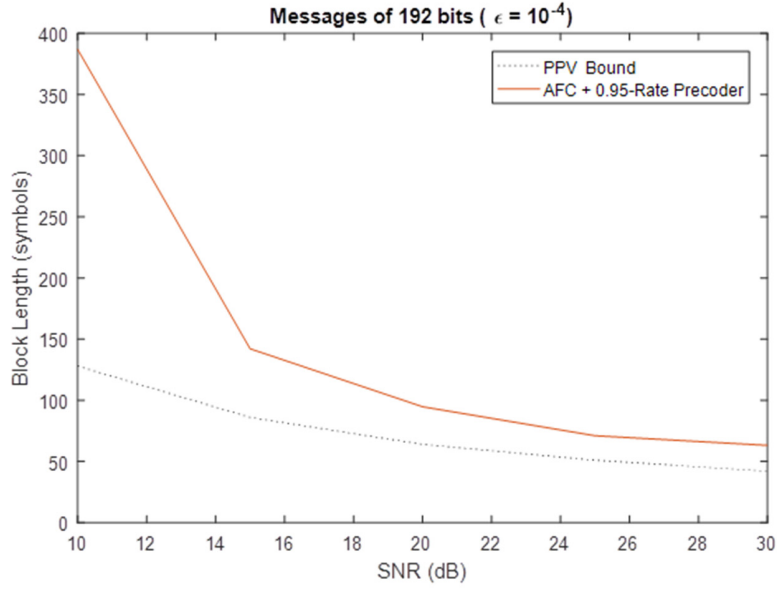

Fig. 2. AFC with a 0.95-rate Protograph-based LDPC precoder are used to encode a message of length 192 bits for a block error rate of 10^{-4} over a wide range of SNRs for the AWGN channel.

Recently, an analog fountain code (AFC) [11] was proposed as a capacity-approaching rateless code over a wide range of SNRs for asymptotically long codewords. AFC can be represented by a single sparse non-binary generator matrix such that the optimization of the coding and modulation can be performed jointly via specialized EXIT charts. The resulting performance is seamless over a large range of SNRs with only linear encoding and decoding complexity with respect to the block length. In Fig. 2, we show that AFC, even in the current sub-optimal design for short codes, has a small gap to the PPV bound in the high SNR region. Moreover, we expect that a much lower latency can be achieved when optimizing AFC for shorter block lengths. As

self-adaptive codes do not require any CSI to be available at the transmitter side, the channel estimation overhead can be eliminated, which has been reported to require 7–8ms in the current LTE standards. Finally, for the sake of completion, it is worth mentioning that our simulations over the Rayleigh fading channel showed that AFC can achieve BLERs as low as 10^{-6} for a wide range of SNRs with space diversity with only 10 antennas and maximum ratio combining.

V. Ultra-fast Signal Processing

The current LTE systems use system throughput as the main design target and performance indicator. In contrast, signal processing latency issues has drawn far less attention in the design process. Similar to Section III, valuable insights into the processing latency bottleneck in the current LTE systems could be obtained by a breakdown of latencies contributed by each LTE receiver module. To this end, we investigate the average computational time for the major receiver modules of an LTE Release 8 system by implementing it on an Intel Core i5 computer. The computational time, a practical indicator for relative latency, is presented in Table II for three typical bandwidths. In the simulations, we have 4 transmit and 2 receive antennas, 16-QAM, and 0.3691 code rate at signal-to-noise ratio of 10dB. The closed-loop spatial multiplexing mode was implemented and the average computational time is based on one subframe. It is clearly shown that MMSE-based channel estimation, MMSE-SIC-based MIMO detection, and Turbo decoding consume the most computational resources and dominate the computational time. To lower the processing latency, new ultra-fast signal processing techniques, especially for the three identified functions, should be developed to strike a favorable tradeoff between throughput and latency.

Receive Modules	B = 1.4MHz	B = 5MHz	B = 10MHz
CFO Compensation	0.0010	0.0023	0.0037
FFT	2.9004e-04	6.2917e-04	8.3004e-04
Disassemble Reference Signal	1.2523e-04	2.2708e-04	3.1685e-04
Channel Estimation (MMSE)	0.0015	0.0141	0.0878
Disassemble Symbols	0.0013	0.0045	0.0087
MIMO Detection (MMSE-SIC)	0.0028	0.0242	0.0760
SINR Calculation	2.4947e-04	6.6754e-04	0.0012
Layer Demapping	4.3253e-05	1.0988e-04	3.8987e-04
Turbo Decoding	0.0129	0.0498	0.1048
Obtained Throughput	2.2739Mbps	10.073Mbps	20.41Mbps

In our simulation, we propose and implement an improved channel estimation approach to reduce the channel estimation latency. The basic idea is to use the least square estimation to extract the CSI associated with the reference symbols, and then employ an advanced low-complexity 2-D biharmonic interpolation method to obtain the CSI for the entire resource block. Typically, the resulting curves from the biharmonic interpolation method are much smoother than the linear and nearest neighbor methods. Our simulation results show that the proposed channel estimation method can reduce around 60% of the computational time relative to the MMSE-based method at B = 5MHz, while achieving almost the same system throughput.

It is also desirable to develop ultra-fast multilayer interference suppression technologies to enable fast MIMO detection, especially for a large number of transmit and receive antennas. Along this direction, a parallel interference cancellation (PIC) with decision statistical combining (DSC) detection algorithm was developed in [12], which can significantly reduce the detection

latency compared with MMSE-SIC. The PIC detectors are equivalent to a bank of matched filters, which avoid the time-consuming MMSE matrix inversion. A very small number of iterations between the decoder and the matched filter are added to achieve the performance of MMSE receivers. This algorithm was also applied to ICI cancellation for high-mobility MIMO-OFDM systems and was shown to achieve a very good performance/complexity tradeoff.

Parallel hardware implementation is another important measure to reduce signal processing latency. For example, the recently proposed parallel turbo decoder architecture [13] eliminates the serial data dependencies, realizes full parallel processing, offers a significantly higher processing throughput, and finally achieves a 50% hardware resource reduction compared with the original architecture. With uRLLC recently declared as one of the major goals in 5G networks, we envisage more research activities in developing ultra-fast signal processing techniques and architectures.

VI. Radio Resource Management

In this section, we will discuss two radio resource management techniques that have great potential to reduce the latency caused by the medium access process.

A. Non-orthogonal Multiple Access

As shown in Table I, grant acquisition and random access procedures in current standards are two major sources of delay. This calls for novel approaches and fundamental shifts from current protocols and standards originally designed for human communication to meet the requirements for ultra-low latency applications. Though optimal in terms of per user achievable rate, orthogonal multiple access (OMA) techniques, such as OFDMA in current LTE, are major causes of the latency associated with the link establishment and random access. More

specifically, in existing wireless systems, radio resources are orthogonally allocated to the users to deliver their messages. This requires the base station to first identify the users through contention-based random access. This strategy suffers from severe collisions and high latencies when the number of users increases.

Non-orthogonal multiple access (NOMA) has recently gained considerable attention as an effective alternative to conventional OMA. In general, NOMA allows the signals from various users to overlap by exploiting power, code or interleaver pattern at the expense of receiver complexity. In the power-domain NOMA, which has been shown to be optimal in terms of spectral efficiency [14], signals from multiple users are superimposed and successive interference cancellation (SIC) is used at the receiver to decode the messages. Users do not need to be identified at the base station beforehand, thus eliminating random access delay which is significantly high in medium to high load scenarios [14].

Fig. 3 shows a comparison between NOMA and OMA in an uncoordinated scenario, where the devices randomly choose a subband for their transmission. The number of subbands is denoted by N_s and the total available bandwidth is assumed to be W = 100MHz. The bandwidth is assumed to be uniformly divided into N_s subbands, each of W/N_s bandwidth. As can be seen, when the number of devices is small, OMA slightly outperforms NOMA in terms of delay, which is expected as the collision probability in this case is small and the devices can achieve higher spectral efficiency as they are transmitting orthogonally. However, when the number of devices is large, NOMA outperforms OMA, as it can effectively exploit the interference and enable the devices to be decoded at the base station. In other words, in high traffic load scenarios, OMA is mainly dominated by the random access collision which leads to unavoidable high

latencies, while NOMA supports a large number of devices with the desired latency, by eliminating the random access phase and enabling the users to share the same radio resources.
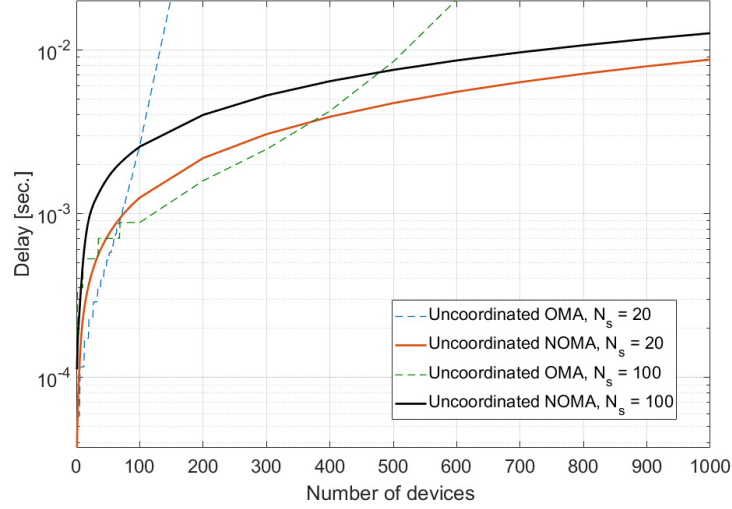

Fig. 3. Delay versus the number of devices for NOMA and OMA.

The main benefits of NOMA come from the fact that it does not need separate grant acquisition and random access phase, as the devices can send their data whenever they want to send. This becomes more beneficial when the number of devices grows large, which is the scenario of interest for most internet-of-things use cases. NOMA can also be easily combined with AFC codes [11] to improve the spectral efficiency for each user, therefore providing a cross-layer solution for reducing the delay. One solution to better satisfy the latency requirements for different applications is to further divide the radio resources between the different uRLLC applications. This will be further discussed in the next subsection. In this way, NOMA can be further tuned to service a larger number of devices with the same requirements.

B. Resource Reservation via Resource Block Slicing

In the current LTE network, the management of radio resource blocks (RBs) for multiple services is jointly optimized. As such, the latencies of different services are interdependent [15].

A traffic overload generated by one service can negatively impact the latency performance of other services. To address this issue, we propose to reserve radio resources for each service. The reservation is done by slicing RBs and allocating a slice to each service based on the traffic demand. Moreover, if RBs in a slice are not used, they will be shared by other services. This type of resource reservation method can achieve a high spectral efficiency and eliminate the latency problem caused by the traffic overload issues coming from other services.

To evaluate the benefit of the proposed RB slicing on a LTE network, we conduct a simulation to compare its performance with a legacy LTE network by using NS-3. Two types of services with different data rates and latency requirements, i.e., low latency intelligent transportation systems (ITS) with average packet sizes of 100 bytes and average packet intervals of 100ms per user, and smart grid (SG) with average packet sizes of 300 bytes and average packet intervals of 80ms per user, respectively, are considered in our simulation. The devices for the above services are distributed in 1 km^2 area according to a Poisson Point Process (PPP) with averages of 400 and 600 devices for ITS and SG, respectively, served by 4 LTE base stations, operating with 20MHz bandwidth. The proportion of traffic load for each slice is approximated based on the ratio of the number of users in a service over the total number of users in all services. Thus, for the proposed RB slicing, we allocate 40% of available RBs exclusively for ITS devices transmissions, leaving the remaining 60% RBs for SG devices transmissions. Note that all available RBs are shared by ITS and SG equally in the current LTE network.

Fig. 4 shows the cumulative density function (CDF) for the end to end packet latencies under a legacy LTE network and under the RB slicing regime that isolates the traffic demand of intelligent transportation systems (ITS) sensor and smart grids (SG) from each other. By performing RB slicing that reserves resources for each service, the latency is reduced from an

average of 10ms to 5ms and 6ms for ITS and SG devices, respectively, as shown in the small box in Fig. 3. This simulation confirms the benefit of the proposed approach. The open future research challenges can be then on how to dynamically optimize the proportion for the resources reserved by multiple services with varying load as well as heterogeneous reliability and latency requirements.
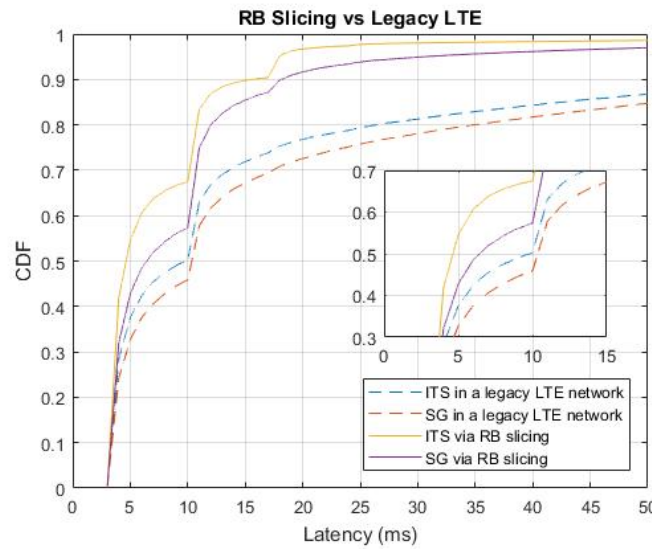

Fig. 4. The cumulative distribution function (CDF) of the end-to-end delay without and with radio resource block slicing[2].

VII. Other Potential Techniques

In addition to the measures introduced in previous sections, there are other techniques that have great potential to reduce the end-to-end latency of cellular systems. In what follows, we briefly discuss the principles of four potential technologies and explain how they can reduce latency.

A. Cross-layer Error Control

Automatic Repeat reQuest (ARQ) is a commonly-used error control method for detecting packet losses by using acknowledgements and timeouts. ARQ has been widely adopted in many

---

[2] The authors would like to thank Zhouyou Gu for his assistance in simulating this figure.

communication networks with Transport Control Protocol (TCP). However, it introduces high and unpredictable delays in wireless networks due to the time varying channel and user contention over a common radio link. On the other hand, User Datagram Protocols (UDP), with no ARQ retransmissions and lower overheads than TCP, have been used for delay sensitive applications with no stringent requirements for low error probabilities, such as Voice over Internet Protocol (VoIP), Video on Demand (VoD), Internet Protocol Television (IPTV) etc.

For emerging mission-critical applications over wireless networks, lower overheads are desirable to reduce overall end to end latency. However, in order for UDP to be suitable for uRLLC, its reliability needs to be substantially improved. Research on this has focused on the design of error control schemes with minimal error protection at the physical layer and rateless coding for erasure channels in the application layer. The research problems have been in optimizing the redundancy split between the physical and application layers to have reliable transmission. This approach involves a significant loss in the decoding error performance due to hard decision decoding at the application layer and weak codes at the physical layer. A promising solution to resolve this is to use short AFC codes in both the physical and the network layer and form a concatenated code with soft output decoding at the physical and soft input decoding at the network layer. Furthermore, the decoding of both AFC codes can be highly parallelized for a low decoding delay.

B. Device-to-Device Communication

Device-to-device (D2D) communication refers to a radio technology that enables direct communication between two physically close terminals. D2D has recently been considered as a key solution for ultra-low latency applications, as it provides a direct link between traffic

participants, without going through the network infrastructure. D2D communication is a good fit for vehicle-to-vehicle (V2V) communications to enable real-time safety systems, cooperative collision avoidance and automated overtake. However, it may be not applicable to many other mission-critical services, such as power systems or remote surgery with communication nodes separated at large distances. Due to the global spectrum shortage, D2D links are expected to operate within the same spectrum used by existing infrastructure-based communication systems (e.g., cellular systems). This calls for highly efficient interference management techniques to ensure the harmonious coexistence between D2D links and conventional links. Otherwise, the latency gain introduced by D2D communication can easily disappear.

C. Mobile Edge Computing

Mobile edge computing (MEC) is a promising approach to promptly process computationally intensive jobs offloaded from mobile devices. Edge computing modules can be installed at base stations which are closer to sensing devices than data servers/clouds. To decrease job-processing delays, edge computing modules are operated in a Software as a Service (SaaS) fashion. In other words, a set of data processing software is in an always-on status, ready to process offloaded jobs from sensing devices. The offloaded jobs can be processed immediately without waiting for computing resource allocation, software initiation, and environment parameter configuration. The data transfer between the sensing device and the computing module in the base station relies on the existing air interface. A multiplexer/de-multiplexer at the base station can distinguish if transmitted data are for computation offloading purpose. If so, the data is redirected to the edge computing modules instead of the mobile core network. In fact, the implementation of edge computing technologies is not mature in cellular networks. The key barrier stems from the incompatibility of computing services and the existing LTE protocol stack. Modifying the

existing stack to accommodate computing services may cause substantial network reconstruction and reconfiguration. Therefore, smoothly merging edge computing into the protocol stack is a key future research direction.

D. Mobile Caching for Content Delivery

Smart mobile caching schemes are also effective solutions for improving the delay performance of data intensive applications, e.g., multimedia, augment reality (AR) applications etc. Mobile caching enables content reuse, which leads to drastic delay reductions and backhaul efficiency improvements. The mobile cache can be installed at each base station. Whenever a mobile device's request "hits" a cached content, the base station intercepts the request and directly returns the cached content without resorting to a remote server. Each base station determines the cached contents through learning their popularities. Caching policies such as geo-based caching and least frequently used eviction, etc. can be employed. The selected contents are then downloaded from remote servers. Downloading cached files is not a delay-sensitive task; hence, it can be operated in a separate network without competing for network bandwidth with other delay-sensitive data traffic. Despite the potential benefits of caching, it is still challenging to realize these benefits in practice. This is because the cache size at the base station is limited, but the number of possible contents can be unlimited. Thus, it is essential to determine how to wisely cache a set of popular contents to maximize the hit rate.

VIII. Summary

This article has introduced the emerging applications, design challenges, and potential approaches in the design of ultra-reliable low latency communications (uRLLC). We described potential use cases of uRLLC in tele-surgery, smart transportation and industry automation and

presented the latency and reliability requirements for these applications. To pinpoint major latency bottlenecks in current cellular networks, we showed a breakdown of the various delay sources in an LTE system and found that a few orders of end-to-end latency reduction is required to support the mission critical applications. To achieve this, each latency component needs to be reduced significantly. Our initial results showed that short analog fountain codes, ultra-fast signal processing, non-orthogonal multiple access and resource reservation via resource block slicing are essential to reduce latency in the physical and multiple access layers. Furthermore, other potential latency reduction measures, including cross-layer error control, device-to-device communication, mobile edge computing and mobile caching, were briefly discussed. We hope this article can encourage more research efforts toward the realization of uRLLC.

REFERENCES

[1] https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D8.4_v1.pdf

[2] N. Larson et al., Investigating Excessive delays in mobile broadband networks, Proc. ACM Sigcom'2015.

[3] Nokia white paper on 5G for Mission Critical Communication,

http://www.hit.bme.hu/~jakab/edu/litr/5G/Nokia_5G_for_Mission_Critical_Communication_White_Paper.pdf

[4] 5GPPP Association, "5G and e-health," 5GPPP, White Paper, Oct. 2015. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf

[5] 5GPPP Association, ``5G automotive vision," 5GPPP, White Paper, Oct. 2015. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf

[6] M. Luvisotto, Z. Pang and D. Dzung, "Ultra High Performance Wireless Control for Critical Applications: Challenges and Directions," IEEE Transactions on Industrial Informatics, vol. 13, no. 3, pp. 1448-1459, June 2017.

[7] Study on Latency Reduction Techniques for LTE (Release 14), document TR 36.881, 3rd Generation Partnership Project (3GPP), 2016. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36_series/36.881/.

[8] P. Schulz et al., "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," IEEE Communications Magazine, vol. 55, no. 2, pp. 70-78, February 2017.

[9] Y. Polyanskiy, H. V. Poor and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2307-2359, May 2010.

[10] G. Durisi, T. Koch and P. Popovski, "Toward Massive, Ultrareliable, and Low-Latency Wireless Communication With Short Packets," Proceedings of the IEEE, vol. 104, no. 9, pp. 1711-1726, Sept. 2016.

[11] Shirvanimoghaddam, Mahyar, Yonghui Li, and Branka Vucetic. "Near-capacity adaptive analog fountain codes for wireless channels." IEEE Communications Letters 17.12 (2013): 2241-2244.

[12] N. Aboutorab, W. Hardjawana, and B. Vucetic, "A new iterative doppler-assisted channel estimation joint with parallel ICI cancellation for high-mobility MIMO-OFDM systems," IEEE Trans. Veh. Technol., vol. 61, no. 4, pp. 1577–1589, May 2012.

[13] A. Li, P. Hailes, R. G. Maunder, B. M. Al-Hashimi, and L. Hanzo, "1.5 Gbit/s FPGA implementation of a fully-parallel turbo decoder designed for mission-critical machine-type communication applications," IEEE Access, vol. 4, pp. 5452–5473, Aug. 2016.

[14] M. Shirvanimoghaddam, M. Dohler, and S. Johnson. "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," IEEE Communications Magazine, Accepted December 2016.

[15] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On radio access network slicing from a radio resource management perspective," IEEE Wireless Communications, vol. PP, no. 99, pp. 2–10, 2017.