This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives

Gordon J. Sutton, Jie Zeng, *Senior Member, IEEE,* Ren Ping Liu, *Senior Member, IEEE,* Wei Ni, *Senior Member, IEEE,* Diep N. Nguyen, Beeshanga A. Jayawickrama, Xiaojing Huang, *Senior Member, IEEE,* Mehran Abolhasan, *Senior Member, IEEE,* Zhang Zhang, Eryk Dutkiewicz, *Senior Member, IEEE* and Tiejun Lv, *Senior Member, IEEE*

***Abstract*—Future 5th generation (5G) networks are expected to enable three key services - enhanced mobile broadband (eMBB), massive machine type communications (mMTC) and ultra-reliable and low latency communications (URLLC). As per the 3rd generation partnership project (3GPP) URLLC requirements, it is expected that the reliability of one transmission of a 32 byte packet will be at least 99.999% and the latency will be at most 1 ms. This unprecedented level of reliability and latency will yield various new applications such as smart grids, industrial automation and intelligent transport systems. In this survey we present potential future URLLC applications, and summarize the corresponding reliability and latency requirements. We provide a comprehensive discussion on physical (PHY) and medium access control (MAC) layer techniques that enable URLLC, addressing both licensed and unlicensed bands. The paper evaluates the relevant PHY and MAC techniques for their ability to improve the reliability and reduce the latency. We identify that enabling long-term evolution (LTE) to coexist in the unlicensed spectrum is also a potential enabler of URLLC in the unlicensed band, and provide numerical evaluations. Lastly, the paper discusses the potential future research directions and challenges in achieving the URLLC requirements.**

***Index Terms*—URLLC, reliability, latency, LTE, unlicensed, coexistence.**

I. Introduction

5th generation (5G) networks are expected to enable three key services, concentrating on each service separately so as to achieve enhanced performance in each. The 5G enhanced mobile broadband (eMBB) service aims to significantly increase the user data rate; the 5G massive machine type communications (mMTC) service aspires to realize the Internet of things (IoT) concept by connecting billions of (often low data rate) smart devices; and the 5G ultra-reliable and low latency communications (URLLC) capability is expected to support unprecedented levels of high reliability and low latency communications. In [1], the 3rd generation partnership project (3GPP) outlines the general URLLC reliability requirement for one transmission of a packet as 99.999% (block error rate (BLER) of 10^{-5}) for 32 bytes with a user plane latency of 1 ms. In [2], the authors propose metrics to evaluate reliability in the time domain, such as mean time to failure, mean time between failures and mean time to repair. For space-domain reliability evaluation, [2] proposes metrics such as the mean covered area and the mean uncovered area, modeled using Poisson point processes (PPPs) and Voronoi tessellation.

Regardless of the metric, it is certain that the unprecedented reliability and latency targets of 5G will give rise to various new and exciting applications, which we discuss in the next subsection.

This paper provides a comprehensive survey on the state of art for URLLC from physical (PHY) and medium access control (MAC) layer perspectives, covering both licensed and unlicensed spectra below 6 GHz. Utilizing the unlicensed spectrum as part of URLLC has not been given much attention previously. To the best of our knowledge, our work is the first which covers URLLC in such depth and provides a valuable insight for researchers who are aiming to work in this area.

A. URLLC Potential Applications

There are a number of potential URLLC applications, which might be operated in either licensed or unlicensed bands, or both, as depicted in Fig. 1. The major use cases include:

- Smart grid [3]: an electrical grid that consists of several operational and energy modules, such as smart meters and devices, as well as renewable energy and energy efficient resources.
- Professional audio [3]: an audio system which is set up by professional live event supporting audio engineers, adopting audio mixers or sound reinforcement systems to perform sound recording, studio music production, sound reinforcement system setup and mixing.
- Self-driving car [4]–[6]: a car which can detect the environment and automatically drive without being operated by a person.

G.J. Sutton, R.P. Liu, D.N. Nguyen, B.A. Jayawickrama, X. Huang, M. Ablohasan, and E. Dutkiewicz are with the Global Big Data Technologies Centre, University of Technology Sydney, Australia (e-mail: {gordon.sutton, renping.liu, diep.nguyen, beeshanga.jayawickrama, mehran.abolhasan, xiaojing.huang, eryk.dutkiewicz}@uts.edu.au).

J. Zeng, the corresponding author, is with Beijing University of Posts and Telecommunications, China, Tsinghua University, China, and University of Technology Sydney, Australia (e-mail: zengjie@tsinghua.edu.cn).

W. Ni is with Data61, Commonwealth Scientific and Industrial Research, Australia (e-mail: wei.ni@data61.csiro.au).

Z. Zhang is with Huawei Technologies Co. Ltd., China (e-mail: zhangzhang4@huawei.com).

T. Lv is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China (e-mail: lvtiejun@bupt.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

- Industrial automation [3]: the new unmanned industrialization beyond mechanization, which processes with the help of control systems, e.g., robots, computers, and information technologies.
- Process automation [3]: an automatic monitoring and decision system for industrial components and procedures, such as heating, mixing, and pumping.
- E-health [7]–[10]: a new healthcare approach with the support of information and communication technology.
- Augmented reality [11]–[13], [15]: a technique to augment the vision of real-world environment by computer-generated information, such as audio, video, and geographic information.
- Intelligent transport system (ITS) [3], [16]: a traffic management system with information and communication technologies, supporting communication interfaces between the elements of road transport, such as vehicles, users and infrastructures, as well as interfaces between different modes of transport.
- Vehicle-to-vehicle (V2V) [17]–[19]: a wireless ad-hoc network which supports communications between vehicles.
- Tactile Internet [9], [20]–[22]: an Internet network, which ensures the tactile sensing with the support of short transit, low latency, high reliability, high availability and high security communications.

Requirements for these use cases are compared in Table I.

B. Recent Advances and Standardization

Challenges, solutions and applications of URLLC can be found in recent literature. At the same time, industry specifications on URLLC are modified and released by standardization organizations.

5G is being standardized in the form of two radio technology components: a novel radio interface denoted as new radio (NR), and long-term evolution (LTE). Achievable latency bounds are evaluated and the expected spectral efficiency is demonstrated in [23]. It is shown that both NR and LTE can fulfil the requirements of international telecommunication union (ITU) 5G. In order to enable low-latency communications, new short slot structures enable faster uplink (UL) and downlink (DL) transmission for URLLC, called mini-slot for NR and short transmission time interval (TTI) for the LTE radio interfaces. In addition, mechanisms to increase the reliability of URLLC services, such as robust coding and modulation, and various diversity schemes, are being developed in accordance with the LTE and NR designs.

The effective bandwidth and effective capacity theories are used in [24] as an analytical framework for calculating the maximum achievable rate for given latency and reliability constraints. The authors point out that the use of a shorter subframe duration for a reduced hybrid automatic repeat request (HARQ) transmission delay could reduce the latency.

A fundamental mechanism is proposed in [25] to revise the methods for encoding control information and data. By combining the header and data of short packets, the combined packet can be efficiently coded, so that the data is delivered faster and with less error. All users have to decode the combined packet, so energy efficiency is traded for very high reliability. [25] also catergorizes ultra-reliable communications (URC) over two dimensions. The first dimension is the time frame used to measure the reliability of the packet transmission (long-term URC and short-term URC). The second dimension is the type of impairment that can affect the communication reliability in a given scenario. Five reliability impairments are summarized, namely, decreased power of the useful signal, uncontrollable interference, resource depletion, protocol reliability mismatch, and equipment failure.

A number of technology components are identified by the mobile and wireless communications enablers for the twenty-twenty information society (METIS) project to address the URLLC requirements, such as reliable service composition framework and operational device-to-device (D2D) links, radio resource management (RRM), MAC, and PHY layer challenges [26]. The trade-offs between bandwith, coding schemes, diversity order, signal-to-noise ratio (SNR) and error rates, when transmitting a 100 bit packet with end-to-end delay of 100 μs, are explored in [27]. The exploration demonstrates that it is feasible to achieve low latency with high reliability by using short transmission intervals without retransmission and equipping base stations (BSs) with a sufficiently large number of antennas to guarantee reliability via a spatial diversity gain.

There are several 3GPP technical reports related to URLLC [1], [28]–[32]. The results of these studies are yet to be standardized but are expected to be included in Release 16 NR technical specifications.

3GPP requirements of URLLC are described in detail in [1], where user plane latency is defined as, "the time it takes to successfully deliver an application layer packet/message from the radio protocol layer 2/3 service data unit (SDU) ingress point to the radio protocol layer 2/3 SDU egress point via the radio interface in both the UL and DL directions, where neither device nor base station reception is restricted by discontinuous reception (DRX)." Reliability can be evaluated as the success probability of transmitting a specified number of bytes within a certain user-plane latency, given a certain channel quality (e.g., a few meters, or coverage-edge). For URLLC, the limit for user plane latency is 0.5 ms for UL and DL separately, and a general requirement for reliability is 99.999% for the transmission of a 32-byte packet with 1 ms user-plane latency.

Put more mathematically, reliability is the probability of transmitting X bytes within a certain end-to-end delay, T, where the end points are the protocol layer 2/3 SDU ingress and egress points. The end-to-end delay may be over one link (e.g., a sidelink), or over two links (e.g., between two user equipment (UE) via the BS. When the transmission is over a single link, the reliability effectively equals 1 - BLER.

The bit error rate (BER) required to meet a given reliability depends on how correlated the decoded bit errors are, which depends on the error correction scheme. To correct for long error bursts that can occur in deep fading or due to bursty interferences, an interleaver which spreads out bursts of errors over time is typically combined with AWGN channel codes [33]. In addition to transmitting the data bits, low-density parity-check (LDPC) codes and turbo codes also transmit

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE I
USE CASE REQUIREMENTS

Use case	Latency (ms)	Reliability (%)	Data Size (bytes)	Communication Range (m)
Smart grid	3 ~20 [3]	99.999 [3]	80 ~1000 [3]	10 ~1000 [3]
Professional audio	2 [3]	99.99999 [3]	3 ~1000 [3]	100 [3]
Self-driving car	1 [4]	99 [5]	144 [6]	400 [6]
Industrial automation	0.25 ~10 [3]	99.9999999 [3]	10 ~300 [3]	50 ~100 [3]
Process automation	50 ~100 [3]	99.99 [3]	40 ~100 [3]	100 ~500 [3]
E-health	30 [7]	99.999 [7]	28 ~1400 [8]	300 ~500 [9]
Augmented reality	0.4 ~2 [12]	99.999 [13]	12k ~16k [14]	100 ~400 [15]
ITS	10 ~100 [3]	99.999 [3]	50 ~200 [16]	300 ~1000 [16]
V2V	5 [18]	99.999 [18]	1600 [18]	300 [19]
Tactile internet	1 [9]	99.99999 [21]	250 [20]	100000 [22]

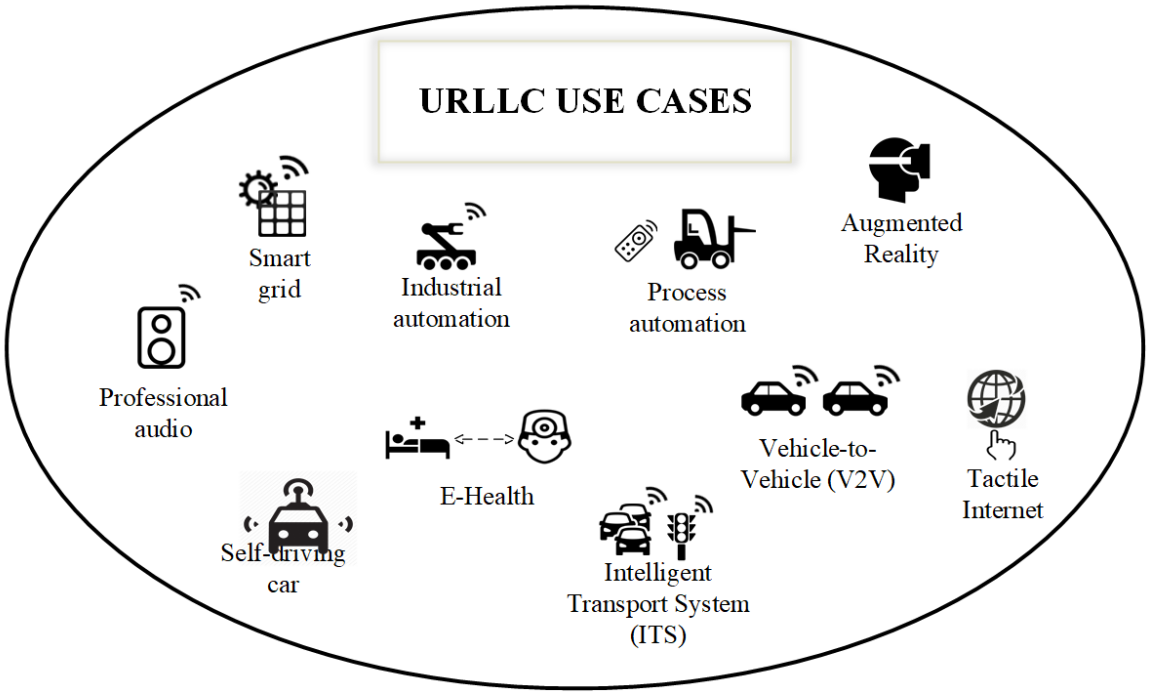


Fig. 1. Potential URLLC use cases

parity bits that equal the parity over a spread of data bits, thereby spreading out bursts of errors and allowing them to be corrected. In the asymptotic case, the Shannon capacity can be approached, with very long code blocks. However, in URLLC scenarios, the block size is typically too short for the codes to effectively ensure reliable communications. The obtainable capacity when transmitting shorter blocks is an emerging field of research, and the design of short codes for small block size has been increasingly attracting attention. We survey the current research of finite blocklength information theory in Section II.A.3) and the performance of short codes in Section II.B.2).

Over a single link (e.g., just UL), if the decoded bit errors are uncorrelated, the required decoded BER, referred to as the information BER (IBER), is related to the BLER and reliability as: reliability = 1 - BLER = (1 - IBER)^{8X}.

When transmitting X bytes in time t_i and bandwidth B_i, the required coding rate is $R_i = 8X/B_i t_i$ (in bits per channel use). Let $IBER_i(t_i, X)$ and $BLER_i(t_i, X)$ respectively be the IBER and BLER on link i in a quasi-static fading channel. From finite blocklength information theory [34]–[36], in typical URLLC transmissions with finite blocklength M, we have [35, eq.3]

$$BLER_i(t_i, X) = \mathbf{E}\left[\epsilon_i(\gamma_i)\right] \approx \mathbf{E}\left[Q\left(\sqrt{\frac{M}{V(\gamma_i)}}\left(C(\gamma_i) - R_i\right)\right)\right], \quad (1)$$

where γ_i is a random variable and denotes the received signal-to-interference-plus-noise ratio (SINR) on link i; $\epsilon_i(\gamma_i)$ is the BLER under a given received SINR γ_i on link i; $\mathbf{E}[.]$ is the expectation function; $C(\gamma) = \log_2(1+\gamma)$; $Q(w) = \int_w^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$, and $V(\gamma) = \frac{\gamma(\gamma+2)}{(1+\gamma)^2} \log_2^2 e$. It is reasonable to assume that the received SINR γ_i is static during one URLLC transmission, since the transmission time can be shorter than the coherence time of the channel [34].

With the decoded bit errors assumed uncorrelated, we have

$$IBER_i(t_i, X) = 1 - [1 - BLER_i(t_i, X)]^{1/(8X)}. \quad (2)$$

The reliability of transmitting X bytes, in time T, over L sequential links, denoted $R(X, T, L)$, can be written as

$$R(X, T, L) = \prod_{i=1}^{L} [1 - BLER_i(t_i, X)], \quad (3)$$

where $\sum_{i=1}^{L} t_i = T$.

Numerologies and frame structure related contents are provided in [31], and a study item is approved that contains scenarios, requirements and technology components for the NR access technology and the channel model for frequency spectrum above 6 GHz. To aid URLLC, mini-slots of length 2-6 symbols are supported for subcarrier spacings of up to 60 kHz [32]. To meet strict 5G URLLC requirements, new study items and work items on URLLC will be carried out and reflected in Release 16 and beyond.

Early predictions for 5G were made in [37] from an IEEE technology perspective, addressing each layer of the protocol stack, but predominantly discussing higher-layer aspects. The aurthor of [37] predicted that devices will need to be able to operate on different wireless networks, and the 5G is expected to have a flat network architecture, with much functionality performed at the base stations, to achieve scalability.

Network and radio interface technologies that enable 5G communications are discussed in [38] and [39]. In [39], an emphasis is placed on utilizing new mm-Wave bands (60 GHz), with directional beamforming, massive multi-input multi-output (MIMO), and the associated spatial division multiple access (SDMA) MAC protocol. Research directions for cellular URLLC are explored in [40], and they instead emphasize non-orthogonal multiple access (NOMA) and coding for latency reduction.

5G communications is surveyed in [41], including network architecture and radio interface technologies. They discuss 5G enabling techniques such as NOMA, sparse coding multiple access (SCMA), massive MIMO, relaying and in-band full-duplex, D2D, and mm-Wave, although there is little discussion on latency or reliability. They also discuss adaptive functionality, such as self-organising networks, cognitive radio and green communications.

The IEEE time sensitive networking (TSN) standard covers link-layer operation, and the deterministic networking (DetNet) standard covers the network layer. A survey of ultra-low latency networks is provided in [42] that focuses on TSN and DetNet. The exploration is broken into flow synchronization, management, control and integrity, and covers ultra-low latency techniques across the wireless access, fronthaul (ethernet/optical), backhaul (optical) and core networks. The latency of ethernet networks is modelled in [43] for a ring topology in industrial settings, and the role of the ethernet in providing low latency vehicle-to-infrastructure (V2I) communications is explored in [44].

In [45], a survey with focus on mm-Wave communications is conducted. The high frequencies of mm-Wave communications create new challenges due to high propagation loss, sensitivity to blockages (e.g., 20-30 dB loss from a human), and the need for directed transmissions. To combat blockages, proposals include utilizing wall reflections, static reflectors, two access points (APs), relays, and spatial diversity. Technological challenges include achieving MIMO and in-band full-duplex at mm-Wave frequencies. The 60 GHz mm-Wave band can only operate over short distances, which aids spectral re-use, but also creates a need for coexistence with systems operating at other frequencies that can transmit further.

Another of the three main services to be supported by 5G is mMTC. The IoT falls into the 5G mMTC category and has been surveyed in [46]. The IoT includes autonomous communication of collected data and control messages to/from smart devices that have sensors and possibly actuators. In the IoT context, ultra-low latency is not an objective, and reliability is instead associated with the success rate of packet delivery without particular focus on latency, for which probabilistic checking of data can be used to identify anomalies and act as a safeguard. With the large number of devices expected in IoT networks, scalability is very important. Cloud and fog computing (i.e., cloudlets or edge computing) have been proposed to achieve scalability. Millions of smart IoT devices connect to thousands of Fog gateways, which connect to hundreds of cloud data centers. A significant portion of the data storage and computing services are performed through fog computing, which, by being closer to the end user, reduces latency. URLLC can also be assisted by upper-layer mechanisms, such as C-RAN, mobile edge computing [47]–[50], network slicing, software-defined networking (SDN) [51]–[56] and caching [57].

The current paper surveys URLLC from PHY and MAC layer perspectives, covering both licensed and unlicensed spectra below 6 GHz. Utilizing the unlicensed spectrum as part of URLLC has not been given much attention previously.

C. Influence of Public Safety Networks

The need for public safety networks, with high reliability and high priority, has driven mission critical communications. Desirable features of public safety LTE (PS LTE) are outlined in [58], including both manual and automatic prioritization adaptation, such as geo-fencing, where user priority changes within a geographical area. Public safety networks can be pre-planned for quick activation, short-lived (triggered by an incident), long-lived (such as a festival) or permanent (in a high-crime area). A particular user can change its priority due to circumstances (e.g., police officer: normal $\rightarrow$ tactical assault role $\rightarrow$ normal). Pre-emption provides a clear path for high priority users by knocking other users off the system, if needed.

The European telecommunications standards institute (ETSI) standard for terrestrial trunked radio (TETRA) [59] has been adopted in many countries and uses a dedicated narrow spectrum. Rather than governments having reserved frequencies for emergencies, 3GPP has been developing LTE mission critical communication standards to enable public safety networks since Release 11. As overviewed in [60], 3GPP Release 11 introduces public safety broadband on Band 14;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

Fig. 2. Paper organization and structure

Release 12 introduces proximity services (ProSe), including D2D, direct discovery, and support for broadcast/multicast; Release 13 introduces mission-critical push-to-talk (MCPTT) [61], enhanced ProSe, and support for single-to-many transmissions; and Release 14 introduces mission-critical data (MCData) and mission-critical video (MCVideo) [62].

Public safety networks are based on group communications, with emergency transmissions received by all members of a group. The current 3GPP functional requirements for group communications are outlined in [63]. A number of the functions have the potential to be used for URLLC, such as admission control, transmission priority and interruption mechanisms. Single-cell point-to-multipoint (SC-PTM) is also possible, which uses a common group radio network temporary identifier (RNTI) and is transmitted on the physical DL shared channel (PDSCH), allowing scalability, without using the multicast channel (MCH).

D. Survey Outline

The organization and structure of the survey is depicted in Fig. 2. This survey aims to explore the PHY-layer, MAC-layer, and cross-layer mechanisms that have the potential to enable URLLC. In Section II, PHY layer mechanisms with the potential to enable URLLC are considered, predominantly from an LTE perspective, covering numerology, diversity and resource reuse. Promising mechanisms include shortening the TTI to reduce the round trip time (RTT), altering the waveform to enable faster decoding, and using finite block-length information theory to reduce the bit error rate. Section III considers cross-layer mechanisms, covering automatic repeat request (ARQ)/HARQ, RRM, multi-connectivity and harmonization.

LTE mechanisms with the potential to enable URLLC for the licensed spectrum are considered in Section IV-A, covering prioritizing bearers during random access (RA) prioritization, admission and when scheduling resources, minimizing control signaling for periodic resource allocations, and using D2D communications to reduce the number of links. In Section IV-B, MAC layer mechanisms used by the incumbent technology in the unlicensed spectrum, i.e., Wi-Fi, are explored. The vehicular network use case is considered in Section IV-C, covering dedicated short-range communications (DSRC) protocols in the unlicensed bands and vehicle-to-anything (V2X) communications in the licensed bands, which rely on D2D communications with semi-permanent scheduling (SPS). In Section IV-D, mechanisms to enable LTE to coexist in the unlicensed spectrum are covered, including current protocols. The challenge for LTE coexistence in the unlicensed spectrum is to maintain the advantages provided by the centrally scheduled LTE protocols, while assimilating with the contention-based Wi-Fi protocols.

The impact of the PHY-layer, cross-layer and MAC-layer URLLC enabling technologies is evaluated in Section V. Potential areas of future research are given in Section VI. The survey is concluded in Section VII.

II. LTE PHY Mechanisms for URLLC

There exists a fundamental correlation among three key performance indicators, reliability, latency and throughput, in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

URLLC, as shown in Fig. 3. To meet the strict URLLC constraints, improvements can be made on each of the three dimensions: reducing latency directly; increasing reliability directly; and improving the throughput with resource-reuse, which can be transformed to improvements in low latency and high reliability. Since there are a number of PHY techniques relevant to latency and/or reliability, we can divide the URLLC related PHY techniques into three categories: structure-based, diversity-based and resource-reuse-based techniques. Generally, structure-based techniques try to reduce latency by shortening the TTI and reducing the symbol duration; diversity-based techniques increase reliability by adding diversity and repetition in the time/frequency/space/code/modulation domain; resource-reuse-based techniques can support low-latency and high-reliability indirectly by cognizing and reusing time/frequency/space resources more precisely. Overall, the state of the art of the PHY techniques enabling URLLC is discussed in this section and a brief summary is shown in Table II. Accurate channel state information (CSI) is important to the error probability and capacity of a wireless link, so we end with a subsection on accurate CSI estimation.

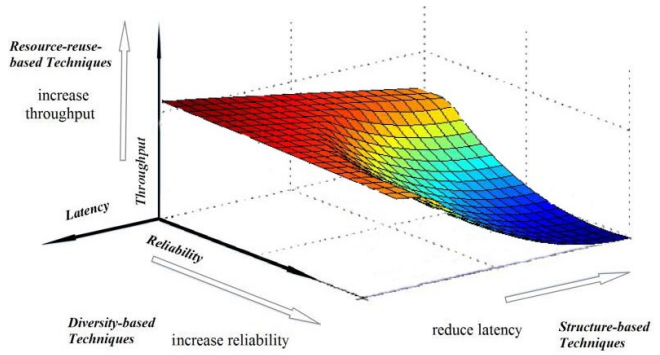


Fig. 3. Using PHY techniques to achieve URLLC from reliability, latency and throughput dimensions.

A. Structure-based Techniques

The legacy LTE numerology is inappropriate for URLLC applications, since it cannot deliver packets within the 0.5 ms user-plane latency requirement of URLLC [1]. Novel designs have been proposed, including the optimization in the number of symbols in one slot, symbol duration, sub-carrier spacing, mapping and modulation methods. These changes directly require improvements in frame structure and waveform design. Meanwhile, the classical Shannon information theory based on the infinite blocklength assumption is limited in the short packet transmission scenario. Thus, the finite blocklength information theory should be studied to support numerology design under short packet and short time duration constraints. Generally speaking, structure-based techniques can decrease the latency sharply without marked loss in reliability. However, these techniques change the framework of LTE system significantly, then facing challenges in hardware implementation and system compatibility, which calls for great evaluation and standardization efforts to be done.

1) Frame Structure: Frame structure is important in the design of 5G PHY aspects, and the ability of achieving lower latency has an intrinsic correlation with frame structure. Numerous studies on URLLC indicate that the design of the frame structure plays an important role in satisfying the 0.5 ms user-plane latency constraint. Especially, shortened TTI, shortened CSI turnaround time, and shorter HARQ have great impacts [64]. Meanwhile, to ensure ultra-reliability, fast processing and wide bandwidth should be considered in the frame structure design. It is also recommended in [64] that URLLC should have specific frame numerologies, or, a highly flexible one to support multiplexing with other use cases, such as eMBB and mMTC, ensuring different number of scheduled users and various types of resource allocations at the same time.

Different design aspects of the control channels (CCH) are analyzed in [65] to support a BLER of 10^{-9}. Then a frame structure is proposed to ensure 1 ms end-to-end user plane latency. Authors of [65] suggest a frame structure that reduces the TTI by a factor of five, from 1 ms to 0.2 ms, and adopts dedicated UL CCHs with a diversity gain to all sporadic traffic users. The simulation results in [65] show that the best choice for the scheduling request (SR) detector might be a coherent matched filter.

A symbol-wise frame with reused numerology, low cyclic-prefix (CP) overhead, and scattered pilot is proposed in [66], achieving higher capacity than the self-contained frame structure, especially at high Doppler scenarios such as V2V and millimeter wave communications. Authors of [67] extend their radio interface design of 5G small cell networks to millimeter wave communications. Furthermore, they discuss potential solutions in the line-of-sight and non-line-of-sight cases separately.

It is pointed out in [68] that the low latency design is mainly supported by the shortened TTI, shorter HARQ, shortened CSI turnaround time, and a faster medium access in UL. The authors propose a 1-symbol based TTI which has ultra-low latency operation and high system capacity.

In [69], the authors propose a flexible frame structure, in which users are multiplexed with separately and dynamically adjusted TTIs. The frame structure is applicable to frequency-division duplexing (FDD), with some features transferable to time-division duplexing (TDD). The frame structure allows for the TTI size to be dynamically adjusted for each scheduling instant of users. Therefore, some users can be scheduled within a short TTI size, e.g., no more than 0.2-0.25 ms, to fulfil the RTT requirement for URLLC. In [70], authors further explore their flexible frame structure with a sufficient user oriented radio resource allocation method, where transmissions have flexible frame sizes.

Considering link level evaluations, system evaluations, and design aspects, it is concluded in [64] that protocol enhancements, reduced processing time, as well as a shortened TTI have potential gains in latency reductions.

The frame structure related techniques are being discussed and evaluated by standardization organizations to determine the most suitable solutions. The limitation of the above proposed frame structures mainly comes from the redesigned nu-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE II
SUMMARY OF PHY TECHNOLOGIES FOR URLLC

Topic	References	Features
Frame structure	[64]–[70]	• Enable low scheduling or processing delay • No obvious increase in process complexity • Shorten TTI and CSI turnaround time • Shorter or no HARQ
Waveform design	[31], [71]–[80]	• Design low-latency waveform • Balance OOBE and reliability
Finite blocklength information theory	[81]–[84]	• Reveal the relationship between reliability and system bandwidth • Low-latency short packet applications • Appropriate bound estimation • Lack of effective bandwidth and capacity theories
Frequency/time/space Diversity	[85]–[90]	• Frequency, time and/or space diversity • Improve reliability greatly • Face deployment constraints
MCS	[91]–[97]	• Obtain diversity from redundancy • Guarantee reliability of short packet transmissions under severe fading • Reduce the latency as well as increase reliability
Frequency hopping	[98]–[102]	• Transmit in different channels successively • Achieve frequency diversity and power gain • No obvious increase in process complexity
Spectrum sensing	[103]–[108]	• Avoid collisions to ensure high reliability • Essential in LBT communications in unlicensed bands
In-band full-duplex	[109]–[114]	• Reduce latency by simultaneously transmitting and sensing • High complexity in SIS
Grant-free NOMA	[115]–[127]	• Reduce latency by grant-free and superposed transmission • Ensure high reliability by advanced receiver design
Accurate CSI estimation	[128]–[131]	• Determine the reliability performance • Grant precise CSI in limited delay budget • Acquire CSI from DL measurement with reciprocity in TDD

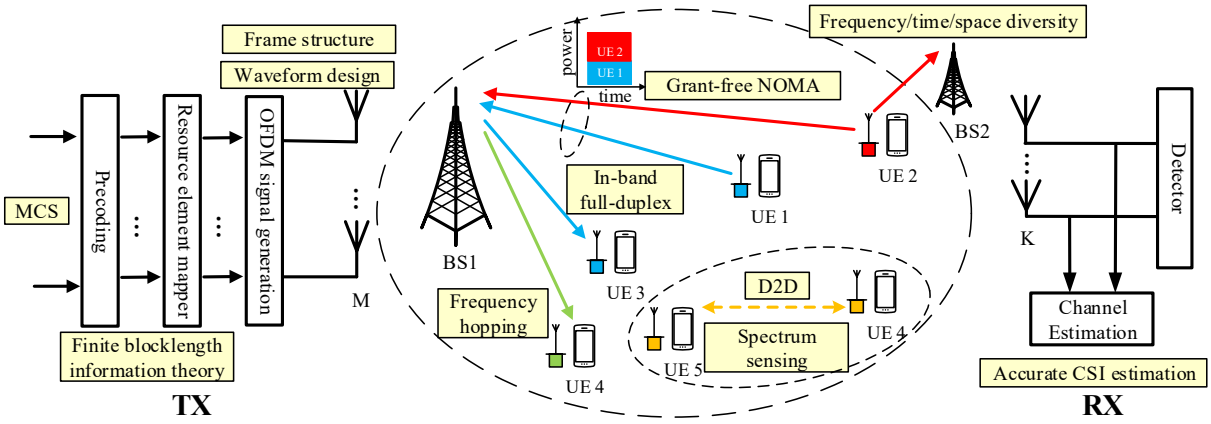


Fig. 4. URLLC-enabling PHY solutions.

merology, from both implementation and compatibility (backwards and forwards) perspectives. It is worth noting that the low-latency frame structure can be ensured by the extremely short time waveform without significant loss in reliability.

2) Waveform Design: Orthogonal frequency division multiplexing (OFDM) is the dominant waveform in 4G and is expected to play an important role in 5G URLLC. OFDM-based waveform is supported in 3GPP Release 14 DL, and the cyclic prefix OFDM (CP-OFDM)/discrete fourier transform-spread-OFDM (DFT-s-OFDM) based waveforms are mandatory for UE in UL [31], [71]. Related techniques for mitigating out-of-band emission (OOBE), such as spectral shaping filtering, time domain windowing, guard band insertion, and spectral precoding are summarized in [72]. A universal framework to balance OOBE and reliability with low complexity is also proposed in [72], based on discrete Fourier transform precoded OFDM.

There are some URLLC related OFDM-based waveforms, such as generalized frequency division multiplexing (GFDM) [73], universal filtered multi-carrier (UFMC) [74] and filtered-OFDM [75]. There are also some DFT-s-OFDM based waveforms, such as generalized DFT-s-OFDM [76] and flexible

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

DFT-s-OFDM [77]. As compared in Table III, potential waveforms in URLLC are suggested to be UFMC and flexible DFT-s-OFDM, considering latency and OOBE regulations which are particularly strict in unlicensed bands.

The authors of [71] compare different filter-based waveforms, including the subcarrier filtering based waveform, the sub-band filtering-based waveform, and the full-band filtering-based waveform. Results show that all filter-based waveforms can significantly reduce OOBE compared to CP-OFDM, thereby being more suitable for unlicensed transmissions. Besides, due to the extra ISI, the filter- based waveforms suffer from reliability degradation. In particular, UFMC has the lowest reliability because the noise is enhanced during demodulation. Meanwhile, the long tail of the impulse response of the filter makes filter bank multi-carrier (FBMC) and GFDM unsuitable for low-latency services. Nevertheless, the other schemes have reduced the length of the impulse response of their filter, which reduces the data transmission and reception times, and allows them to support low latency services.

The authors of [76] investigate whether OFDM, DFT-s-OFDM, FBMC-OQAM, UFMC and G-DFT-s-OFDM support ultra-low latency by using shorter symbols. They find that the latency of FBMC-OQAM is compromised by the need to accommodating long filter tails, and that G-DFT-s-OFDM can be decoded quickly by transmitting only a portion of the IFFT output. In addition, paper [75] shows that windowing OFDM (W-OFDM) achieves limited OOBE improvement. Instead, filtered-OFDM provides more protection against interference from different numerologies compared to W-OFDM, which results in better BLER performance (almost the same as CP-OFDM), higher spectrum efficiency, lower OOBE with sharp transition regions, and lower time-domain overhead.

A software-defined air interface (SDAI) is presented in [78] under a unified framework, in which the frame structure, waveform, multiple access, duplex mode, and antenna configuration can be adaptively configured. A compatible low-complexity multi-carrier modulation structure is also proposed in [78]. A framework that can develop a flexible numerology and waveform is proposed in [79]. Thus, waveform flexibility can be extended in 5G, e.g., large subcarrier spacing is preferable for URLLC applications due to shorter symbol duration. Focusing on efficient generation and processing of subband-filtered CP-OFDM signals, fast-convolution-filtered-OFDM (FC-F-OFDM) is proposed in [80] to allow arbitrary subband configurations to be constructed freely.

However, many URLLC use cases will be implemented in unlicensed bands, which means the OOBE of the new waveforms should be restricted to a tolerant level. Therefore, several power-spectrum density-limiting schemes should be adopted in the OFDM-base waveforms to further decrease the interference with other users and systems. At the same time, the waveform design has to suit the fast-speed wide-band signal processing and the particularity of short packet transmissions.

3) Finite Blocklength Information Theory: Based on an infinite blocklength argument and random coding scheme, the well-known Shannon capacity model noticeably underestimates the delay for finite blocklength packet transmissions, which potentially causes inefficient radio resource allocations. Finite-blocklength information theory reveals the relationship between the desired reliability and system bandwidth in low-latency short packet applications.

Exploiting the stochastic network calculus, authors of [81] compute probabilistic delay bounds for low latency wireless systems in fading channels. Consequently, they provide a service characterization and point out that finite blocklength performance models need to be extended up to the application layer with queuing effects considered.

Recent developments in finite blocklength information theory are presented in [82], where the authors propose bounds on the maximum number of bits that can be transmitted within given bandwidth, latency, and reliability constraints. The bounds unveil the fundamental interplay between latency, bandwidth, rate, and reliability. Authors of [83] investigate the scenarios in MIMO Rayleigh block-fading channels. They calculate the upper and lower bounds of the highest coding rate for finite-blocklength, finite-SNR and specified BLER constraints. These bounds reveal that there is a balance between the rate gain obtained from available degrees of freedom from time, frequency and spatial domains, and the rate loss caused by the estimation of fading coefficients over many domains. In [84], recent advances in the theory of finite blocklength packet transmissions are provided. It is verified that novel communication protocols designed by finite-blocklength information theory are efficient in some typical scenarios.

The study on finite-blocklength information theory is still in its infancy. There is a need to further develop the effective bandwidth/effective capacity theories, which describe the relationship between reliability and latency more logically and with closed-form expressions.

B. Diversity-based Techniques

Reliability is a permanent objective in wireless communications, with requirements becoming more rigorous in URLLC. Diversity, with redundancy, plays a leading role in boosting reliability, so that random noise and errors do not neccessarily lead to packet loss. The most mature technique to improve reliability in erasure and noisy channels is the modulation and coding scheme (MCS). The MCS can be considered as obtaining diversity from redundancy over time (especially for low rate codes, such as repetition code and fountain code). For low latency, the additional coding delay needs to be reduced by parallel-computation design. Spatial diversity, which is obtained from distributed input and output antennas, can keep transmissions almost error-free without particularly increasing the latency. Frequency hopping also increases reliability in the licensed and unlicensed spectrum over frequency-varying fading channels by achieving a frequency diversity gain. Current diversity-based techniques are able to meet URLLC reliability requirements, owing to profound theoretical research and numerous low complexity implementations.

1) Frequency/Time/Space Diversity: Diversity gains can be achieved in the frequency, time and/or space domains, as depicted in Fig. 5. Diversity is widely regarded as a crucial and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE III
WAVEFORMS PROPERTIES COMPARISON

Symbol	Complexity	OOB	Ultra-low Latency
GFDM	High receiver complexity, as a result of large-sized FFT and the SIC used [75].	Adopt filtering and windowing methods to control OOBE [73].	Detection process can only be started after the entire block is received [75].
UFMC	It requires filtering per block of subcarriers [76]. Zero padding (ZP) and 2N FFT is used at the receiver, increasing the complexity [75].	The sub-band filtering in UFMC leads to a reduced OOBE [75].	Enabled by using shorter symbols [71].
filtered-OFDM	Relatively high complexity due to the filtering operation [75].	For non-adjacent resource block (RB), it is hard to reduce the OOB between the RBs with a band-pass filter [71].	–
Generalized DFT-s-OFDM	It requires (DFT +IFFT) / (FFT + IDFT) at transmitter/receiver [76].	–	Enabled by using shorter symbols [76].
flexible DFT-s-OFDM	It requires (DFT +IFFT) / (FFT+ IDFT) transmitter/receiver [76].	–	Enabled by using shorter symbols and/or transmitting only a portion of the IFFT output [76].

efficient way to achieve the URC without an obvious increase in latency.

The use of several less reliable links instead of one dependable link is proposed in [85] to ensure high reliability. With the optimized transmission power allocation over parallel links, the desired reliability can be achieved with energy efficiency. Considering both noise-limited and interference-limited scenarios, the coverage and capacity for a realistic factory setting are analyzed in [86]. The analysis reveals that diversity affects the system capacity markedly. Authors of [87] propose a new data transmission scheme based on spatial diversity with cooperative ARQ and accurate feedback to improve the system performance. Since the proposed scheme relies heavily on accurate feedback, the spatial diversity gain is related to the CSI uncertainty. In [88], the impact of spatial and frequency diversity on reliability and the required bandwidth is studied using a two-state transmission model that adopts finite blocklength channel codes.

Recently, massive MIMO is emerging as the key enabling techniques of 5G, due to the significant improvement in spatial diversity, supporting tens of users in different beam directions simultaneously. In [89], the authors review the reliability and latency performance of the millimeter wave enabled massive MIMO system. Channel variations, system dynamics, and probabilistic constraints on reliability and latency are included in a network utility maximization problem via Lyapunov techniques, achieving 99.99% reliability with significant latency reductions. In [90], the authors study the scenarios in which a few antennas are equipped at the transmitter, e.g., sensor networks. They verify that the large antenna arrays at the receiver, the URC can be guaranteed by coherent or non-coherent receivers, even with a 2×64 antenna configuration. They point out that a sensor with a single or two antennas is able to achieve ultra-high reliability with a massive MIMO BS. In addition, accurate CSI estimation can greatly further improve the spectral efficiency of space diversity, achieving higher transmission beamforming gains, particularly in scenarios with multiple transmitting antennas.

In URLLC applications, the operation spectrum is always predefined with restricted bandwidth, such that the frequency diversity is limited to no more than ten. Also, due to the short transmission time needed to acheive URLLC, the channel does not vary significantly, so time diversity is strictly limited. Hence, the biggest diversity gain comes from space diversity, which can be in the 100's when massive MIMO is adopted. However, the antenna array size might be too large in bands below 6 GHz, bringing challenges to deploying this scheme. Millimeter wave enabled massive MIMO systems, with small component sizes and wider bandwidths, might be a promising research hotspot in URLLC.

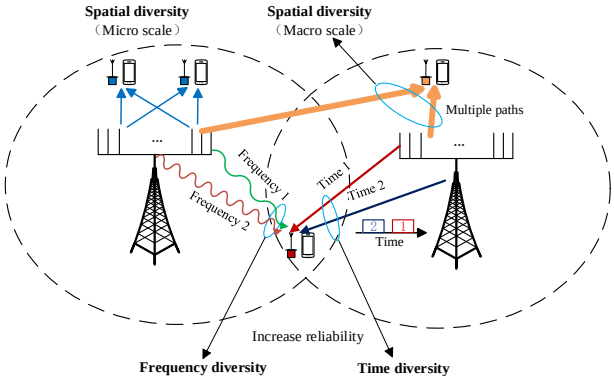


Fig. 5. URLLC using diversity in time, space, and frequency at both micro and macro scales.

2) Modulation and Coding Scheme: Trying to mitigate the effects caused by different channel fading, the MCS is definitely an important part of URC. However, in URLLC scenarios, low latency coding/decoding and modulation/demodulation schemes are urgently needed to meet the strict processing latency requirement.

The authors of [91] propose a constrained convex resource allocation framework suitable for jointly optimizing both the MCS indexes and the code sparsity of random linear network coding (RLNC). In addition, the proposed schemes can ensure a significant reduction in the average number of decoding operations of at least 92% and 57% for ultra-reliable layered multicast communications without altering the actual implementation of the decoder. In [92], the authors derive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

the optimal adaptive modulation and coding (AMC) in URC to achieve the maximum throughput (MT), and propose a sub-optimal limited feedback AMC scheme that obtains the near optimal link adaptation to support URLLC efficiently. Power control and rate adaptation are studied in [93] to minimize the delay for concurrent transmissions of sensor nodes, in which a few transmission rates can be supported. An optimal polynomial time algorithm is then proposed to solve the problem. It is presented in [94] that punctured trellis-coded modulation (TCM) can obtain a high coding rate flexibility and a low decoding complexity, thus being an attractive alternative to MD-TCM for low-latency applications with high spectral efficiency. In [95], a robust link adaptation is enabled to support user data rates with a given reliability in a precoded DL system. Data rates are selected that provide high decoding reliability, while adhering to transmission delay constraints. Simulation results demonstrate that the accuracy of fading characterization and statistics is crucial for robust link adaptation. In [96], some recent codes, e.g., LDPC, extended Bose-Chaudhuri-Hocquenghem (BCH) code, turbo code and tail-biting convolutional code (TB CC), are investigated and the comparison of their performance is conducted under finite-blocklength and low decoding-complexity limitations. From the investigation, we find out that TB CC based on a memory-14 encoder with [75063 56711] generator polynomials outperforms short LDPC, extended BCH and turbo code in BLER over the bi-additive white Gaussian noise (AWGN) channel when the blocklength is 128 bits and the code dimension is 64 bits. However, the decoding complexity of TB CC, caused by the enormous number of possible states of the code trellis, needs to be further reduced.

The analog fountain code (AFC), first proposed in [97], is a rateless code that approaches capacity limitations over a wide range of SNRs. In [40], the authors verify that AFC can be optimized in short codes to achieve 10^{-6} BLER in a Rayleigh fading channel with 10 antennas at the receiver.

Novel modulation and coding methods are still urgently needed to satisfy the short packet features. Especially, fast coding and decoding should be supported.

3) Frequency Hopping: Frequency hopping enables transmissions in separate channels successively in a predefined sequence and can be used to achieve high reliability with low latency when the information needs to be transmitted immediately without precise CSI. As shown in Fig. 6, it can achieve a high frequency diversity gain in rich scattering environments.

Paper [98] proposes a new UL physical layer architecture which consists of subcarrier hopping, super-orthogonal convolutional codes and Golay complementary sequences to attain high coding gain, frequency diversity, and low PAPR. In [99], authors study the super-trellis decoding and successive interference cancellation (SIC) in a subcarrier hopping multiple access system. The authors reveal that, based on proper design, SIC achieves proximate reliability to the optimum super-trellis decoding with much lower complexity for low latency transmissions. In [100], authors propose a shortened physical UL control channel (sPUCCH) which contains two single carrier-frequency division multiple access (SC-FDMA) symbols to improve reliability in ultra-low latency communications by utilizing symbol-level frequency hopping. As presented in [101], a redundant slot-level channel hopping approach can be employed where several frequencies are allocated to every link. In [102], channel hopping algorithms are developed for symmetric-synchronous, symmetric-asynchronous and asymmetric-asynchronous environments with dynamic spectrum sharing.

Furthermore, fast frequency hopping can be combined with the MCS and forward error correction design to allow duplicated bits to be sent together in time, while also having low-correlated noise and fading.

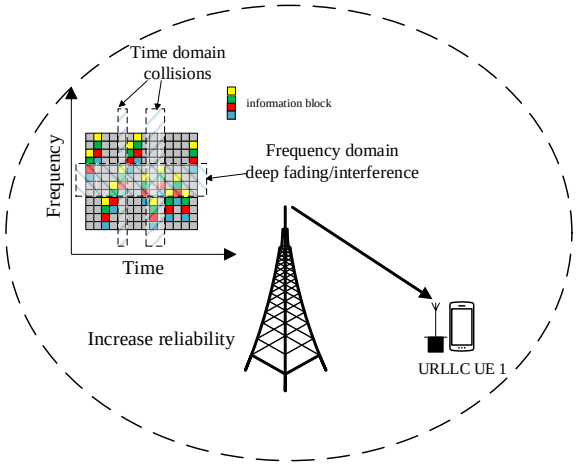


Fig. 6. Frequency hopping in unlicensed channels

C. Resource-Reuse-based Techniques

Different from the structure and diversity based techniques mentioned above, which aim at directly achieving latency and reliability requirements separately, resource-reuse-based techniques can cognize and reuse time-frequency resources more precisely to satisfy URLLC requirements indirectly. Spectrum sensing can locate temporarily free channels and monitor the overloading state of occupied channels accurately. In-band full-duplex nearly doubles the capacity by simultaneously transmitting and receiving/sensing in the same time-frequency resource, ultimately supporting low latency and high reliability. The recently emerging grant-free NOMA supports instant short packet transmissions from different users in the same time-frequency resource without a noticeable decrease in BLER, owing to the advanced receiver. Overall, techniques based on resource reuse benefit from reuse or precise utilization of resources, but a breakthrough in computational complexity reduction is necessary to support further standardization and industrialization.

1) Spectrum Sensing: Spectrum sensing, as depicted in Fig. 7, is essential in listen-before-talk (LBT) communications in unlicensed bands. Currently, low complexity and fast wideband sensing schemes should be supported in URLLC deployment because the sensing time limitation is still rigorous.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

The authors of [103] provide a thorough investigation on the research status, standardization, and applications of spectrum sensing, and then point out the potential challenges and future research directions. In [104], the authors categorize and review the energy efficient algorithms in cooperative spectrum sensing, which are more reliable but more complex than single-device spectrum sensing algorithms.

In [105], authors propose a zero-block sub-Nyquist sampling detection scheme to precisely detect spectrum holes with low complexity. A coordinated multi-channel spectrum sensing (Cluster-CMSS) policy is proposed in [106] to precisely detect channels with maximum empty probability for secondary users. In [107], reliable and energy-efficient detection methods, based on sparse multi-channel signal processing, are proposed for individual and cooperative wideband spectrum sensing scenarios. In [108], a novel individual spectrum sensing method is proposed that exploits historical sensing data to improve the preciseness, and detailed steps of an algorithm to implement the proposed method through Gibbs sampling are given.

The fast wide-band spectrum sensing with high precision is helpful in URLLC. Both individual and cooperative spectrum sensing schemes can also be utilized. Further studies into energy efficient sensing schemes will benefit URLLC, since spectrum sensing always runs in the background in energy-limited URLLC use cases.

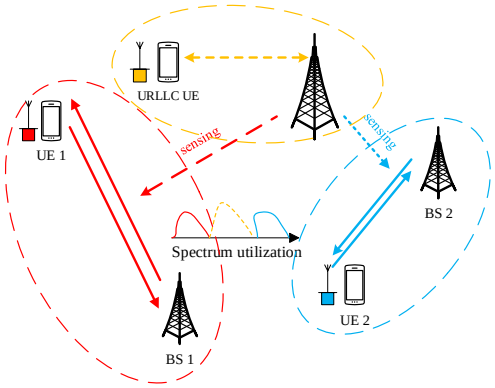


Fig. 7. Spectrum sensing.

2) In-band Full-duplex: Recent advances in signal processing and antenna design allow radios to suppress or cancel their transmission signals in the receiving chains, which is referred to as self-interference suppression (SIS). SIS enables in-band full-duplex communications by allowing radio transceivers to transmit and receive simultaneously on even the same antenna array (see [109] and references therein). SIS can also be utilized in opportunistic spectrum access systems, in which secondary users sense the spectrum continuously and access a free channel opportunistically, as depicted in Fig. 8. Then, a secondary user can mitigate the unwanted interference of its simultaneous transmission signals to lower the sensing delay and the collision probability. Thus, in-band full-duplex has prospective benefits in URLLC deployment, especially in unlicensed bands.

In [110], the authors comprehensively investigate the basic concept, suppression techniques, MAC protocols, and performance of in-band full-duplex systems, and then they indicate the research trends and potential applications. Authors of [111] analyze in-band full-duplex transmitting and sensing, and formulate the relationship between sensing and spectrum awareness. They then propose an optimal adaptive method for use in overlay opportunistic spectrum access systems. In [112], a three-stage switching method is proposed to determine when to sense the spectrum simultaneously. The authors also explore different spectrum sensing schemes with SIS, which can be used for in-band full-duplex operation. [113] studies the performance of spectrum sensing schemes that use either a single-channel or multi-channel energy detector for in-band full-duplex operation. In this scenario, in-phase and quadrature imbalance (IQI) in the joint transmitter-receivers causes a significant increase of the false alarm probability and a noticeable decrease of the detection probability. Thus, IQI needs to be carefully considered in in-band full-duplex operation design. In the energy and traffic aware in-band full-duplex communications, the authors of [114] derive the resource allocation optimization problem under the energy and load constraints, and then they propose a sum-optimal solution to minimizing the data queuing delay of UE from the perspectives of beamformers, scheduling and resource allocation.

In-band full-duplex has a great potential to improve the spectrum and medium awareness through its simultaneous transmitting and receiving/sensing capability. Nonetheless, in-band full-duplex receiving chains still suffer from the leaked self-interference caused by imperfect SIS techniques. This factor has to be taken into account in network planning and power/interference budgeting when applying in-band full-duplex in URLLC.

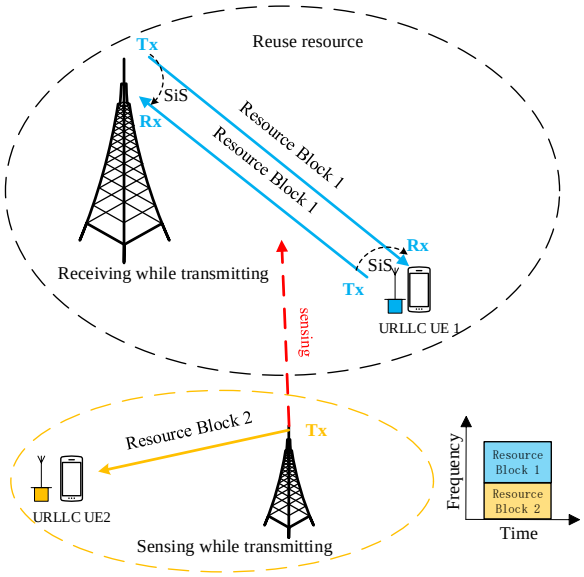


Fig. 8. In-band full-duplex illustration.

3) Grant-Free NOMA: In URLLC, grant-free NOMA, which is depicted in Fig. 9, can remove the grant-request and the scheduling process, thereby significantly reducing the sig-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

naling overheads and air-interface latency without substantially decreasing the reliability.

Scheduling latency can be reduced by implementing grant-free transmissions. A UE can be configured with grant-free resources, allowing it to transmit without scheduling. The time-frequency resources, along with a dedicated reference signal, are pre-configured to the UE semi-statically for URLLC UL grant-free transmission. Frequencies for hopping between initial transmission and re-transmissions should be also provided to reduce repeated collisions.

NOMA is an efficient way to resolve packet collisions by taking the advantage of interference cancellation [115]–[119]. Low-correlation spreading sequences and asynchronous HARQ in the UL can be considered to reduce the collision probability.

In [121], NOMA is regarded as one of the key enabling technologies to fulfil the requirements of 5G. An ultra-dense network is simulated, showing that NOMA with full-duplex can have much higher sum rate than both OMA with half-duplex and NOMA with half-duplex.

A subclass of NOMA is signature-based NOMA (S-NOMA), in which predefined signatures, spreading sequences or scrambling sequences, are generated from device-specific codebook structures. In [120], the authors comprehensively investigate the signature design and multi-user detection (MUD) for S-NOMA. S-NOMA schemes are demonstrated to be sensitive to impulsive noises, through simulation, and several challenges and future research directions in S-NOMA are pointed out.

With NOMA, more than two users can be served with the same time and frequency resource. In [122], a MIMO-NOMA DL transmission scenario is studied to enable one BS equipped with multi-antenna to support two users on the same time-frequency resource with high reliability. In [123], a low complexity threshold-based message passing algorithm (MPA) is proposed to achieve sub-optimal performance when a belief threshold is properly set to rapidly select credible codewords. In [124], a grant-free rateless multiple access scheme is proposed to reduce latency remarkably by utilizing the inherent pseudo-random pattern. In [125], the authors propose a grant-free sparse code multiple access to significantly reduce the transmission latency in heavy overloading scenarios.

Grant-free NOMA is also supported in [40] to avoid RA collisions when the traffic load is heavy, thereby reducing the latency without loss in reliability. Consequently, the authors point out that the natural combination of NOMA and AFC can further reduce the latency, and then provide a total solution for URLLC. In [126], the authors utilize a dynamic compressive sensing (DCS)-based MUD to exploit the temporal correlation of active user sets. In particular, the estimated active user set over the previous time slots is treated as the prior information of the following time slot. In [127], a prior-information-aided adaptive subspace pursuit (PIA-ASP) algorithm, which improves the MUD performance in Grant-free NOMA, is introduced. Then, a robust PIA-ASP algorithm is further proposed to improve the estimation accuracy.

While grant-free NOMA has great potential, MUD in grant-free NOMA is currently computationally complex and requires a large number of iterations. To be deployed widely, processing time needs to be reduced. In addition, the interference becomes more pronounced under a network supporting grant-free NOMA. As such, grant-free NOMA needs to be further studied and more fully understood so that the in-band and out-of-band interference, both within the system and to other systems deployed in the same and adjacent bands, can be managed.

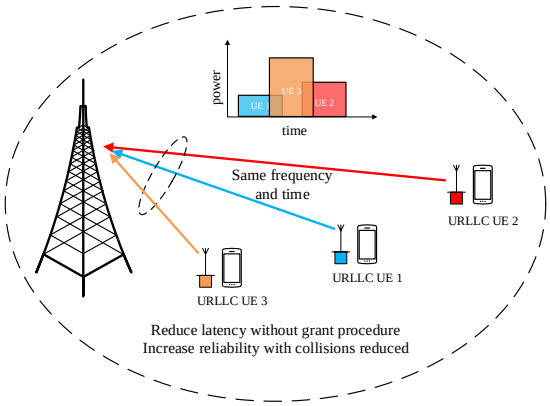


Fig. 9. Grant-free NOMA illustration.

D. Accurate CSI Estimation

CSI measurement is critical in unlicensed URLLC, as precise CSI should be estimated and fed back in a strictly limited delay budget. Further, numerous challenges emerge in CSI quantization with modest information bits, and fast accurate CSI feedback. However, in a TDD system, UL CSI could be acquired from the DL CSI measurement with reciprocity. Then, reliability increases, and interference decreases with transmitting beamforming enabled by the fast CSI measurement at the transmitter.

Estimating the angle-of-arrival (AoA) quickly and accurately is important for accurate parallel CSI estimation [128]–[131]. In [130], the authors propose a low computational complexity AoA estimation by estimating signal parameters via rotational invariance technique (ESPRIT) for the massive MIMO system with two kinds of hybrid subarrays, referred to side-by-side and interleave sub-arrays. The authors of [131] propose an approach that allows subarrays to use different phase shifts per estimation to resolve the ambiguity problem by directly estimating the desired AoA parameters. This approach can speed up the estimation and improve the estimation performance, which is suitable in URC with a processing time limitation.

III. Cross-layer Mechanisms for URLLC

Cross-layer-related URLLC techniques are discussed in this section and are summarized in Table IV. These mechanisms are categorized as cross-layer because they access the data from both the PHY-layer and MAC-layer to exchange information and enable interactions. The HARQ/ARQ mechanisms

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

create automatic error-correction loops between the MAC and PHY layers. RRM, multi-connectivity and harmonization organize resources and/or links, using knowledge of current channel conditions.

A. ARQ/HARQ

ARQ and HARQ are important mechanisms that allow LTE to balance spectrum efficiency and reliability. HARQ lies in the MAC layer and uses forward error correction to automatically trigger an additional transmission, containing extra error-correction bits, when it detects and cannot resolve an error. Errors not detected/resolved by the HARQ process continue to the radio link control (RLC) layer, where the ARQ process uses the cyclic redundancy check (CRC) to detect errors, and automatically trigger a retransmission if needed. Although these schemes can fully exploit time diversity to increase reliability, especially in time-varying channels, it is difficult to support more than one retransmission within a 1 ms latency constraint. Thus, the contributions of ARQ/HARQ to URLCC might be limited. However, cooperative ARQ which sets up a virtual antenna array among BSs is proposed to guarantee reliability under severe shadowing fading.

The reliability performance of a cooperative ARQ system with a short retransmission delay is analyzed in [132] and found to sharply reduce the BLER. The efficiency of HARQ and the impact of CSI feedback accuracy are analyzed in [133]. They recommended that only one retransmission be supported to ensure low latency, and found that precise CSI feedback can improve the reliability. The authors of [87] also explore the influence of CSI feedback accuracy on ARQ performance, and then propose a new spatial diversity-based data transmission scheme with cooperative ARQ to improve the transmission reliability in a smart factory scenario. In [134], the authors propose a new HARQ process pooling method for multi-connectivity UE, in which dynamic and adaptive splitting HARQ processes are cooperatively handled across different carriers to support URC.

The end-to-end reliability for communications between two UE, via a BS, is considered in [135]. For a latency budget of 1 ms, a 0.125 ms TTI is used to allow four transmission attempts, which can occur on either link. Rather than allocating resources equally to the UL and DL, resourcing two transmission attempts on each link, the number of transmission attempts and the UL and DL resources are adjusted based on the instantaneous CSI of both links to achieve a target reliability in terms of BLER. In a fixed assignment scheme, the number of attempts and resources are preconfigured. In an adaptive assignment scheme, the number of attempts and the resources can be altered between transmissions, via acknowledgment (ACK)/negative acknowledgment (NACK) feedback and assuming minimal processing delay, to account for the SINRs over the two links. [135] formulates the optimization problem using ARQ and incremental redundancy hybrid ARQ (IR-HARQ) schemes and offers suboptimal solutions to reduce complexity. Both the fixed and adaptive transmission assignment schemes reduce the required resources compared to the default equal UL/DL resource assignment, when there is a mismatch in the channel quality of the two links.

Time-critical data may not benefit from retransmission, as retransmissions can cause congestion and increase delays. The 5G public private partnership (5G PPP) project METIS-II has flagged the idea of turning off the HARQ function in some scenarios to achieve URLLC [136]. This idea has already been captured in the unlicensed spectrum Wi-Fi protocol, where Wi-Fi quality of service (QoS) stations (STAs) have two service classes, QoSNoAck and QoSAck. Frames from service class QoSNoAck are not acknowledged at the MAC level, and so are not retransmitted [137].

B. Radio Resource Management

RRM aims to ensure spectrum efficiency and energy efficiency, while suppressing inter-cell and intra-cell interference. RRM operates across the PHY/MAC layers and combines channel conditions, obtained via the CSI, with MAC scheduling, to dynamically allocate resource blocks and control transmission power levels. In unlicensed bands, power control also plays an important role in coexistence of different systems while reducing the OOBE. As such, RRM is an important component of URLLC.

After investigating the RRM problem in V2V communications, the authors of [138] formulate QoS requirements. They then propose a separate resource block allocation and power control algorithm to maximize the sum data rate while satisfying latency and reliability requirements. In [139], a queue-state and channel-state information-dependent transmission policy is proposed to achieve energy efficiency, and a multi-user bandwidth allocation method is also studied.

There is a regulatory transmit power control (TPC) requirement in some regions, where devices are required to reduce the maximum transmit power by 3 dB or 6 dB [140]. A novel approach is developed in [141] that utilizes causal knowledge of data arrivals and latency constraints to obtain energy-efficient scheduling for latency restricted finite blocklength communications. A spatially-dynamic open-loop power-control solution is introduced in [142] that controls the transmission power of users, aiming to mitigate the interference between cellular and D2D transmissions. Simulation results demonstrate that the solution can foster the co-existence of cellular and D2D systems.

C. Multi-connectivity

Multi-connectivity means maintaining access through more than one connection, which is commonly used in soft-handover and dual-connectivity hotspots. Nowadays, it is usually mentioned with the concept of C-RAN which enables centralized baseband processing of signals collected from multiple remote radio heads. However, D2D and drone-assisted access are also considered to be other forms of connectivity. Generally speaking, multi-connectivity adopts space diversity to ensure ultra-reliability without a marked increase in latency at the price of the complex cooperation in networking.

[143] presents an user-centric dynamic radio access network (RAN) selection method and traffic load adaptation,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE IV
SUMMARY OF CROSS-LAYER TECHNOLOGIES FOR URLLC

Topic	References	Features
ARQ/HARQ	[87], [132]–[137]	• Affects the latency by multiple retransmissions • Omitted HARQ, for time-critical data, to reduce congestion and latency • Adaptive retransmissions, to improve dual-link reliability for same latency
RRM	[138]–[142]	• Accurate CSI are needed to achieve high reliability, spectrum and efficiency • TPC increases energy efficiency and reduces OOBE
Multi-connectivity	[143]–[147]	• Enable centralized baseband processing • Integration of cellular, D2D and drone-assisted based connectivity • Increase reliability by spatial diversity and fading avoidance • Increase networking cooperation complexity to avoid high latency
Harmonization	[136], [148]	• Combines functionality of multiple protocol stacks into one, at and above harmonization layer • At PDCP layer: easy coordination of different air interfaces (e.g., LTE and NR) • At MAC layer: flexible scheduling, accommodating different MAC schemes, algorithms, parameters (e.g., contention-based, scheduled, RTS/CTS, LBT, and prioritizations)

thereby achieving energy-efficient high quality health monitoring in heterogeneous network. [144] proposes a multi-connectivity method in the C-RAN to reduce mobility related link failures and improve the cell-edge throughput.

[145] revisits prior art schemes for managing the set of co-ordinating cells, referred to as the active set (AS), and compares them against a new proposed AS management scheme. It is shown that with a fixed AS size multi-connectivity scheme, ultra-reliability in terms of resolving of radio link failures (RLFs) is guaranteed. In [146], the authors propose a C-RAN and D2D combined architecture to handle the associated fronthaul delay of C-RAN. They believe that this architecture can be deployed in licensed and unlicensed bands with almost "zero delay". The availability of alternative connectivity options, such as D2D links, cellular connectivity and drone-assisted access is discussed in [147]. D2D connections and drone-assisted links are highly utilized to improve the availability and reliability of URLLC data acquisition for devices moving at low and moderate speeds.

D. Harmonization

Harmonization is the combining of potentially different specialized solutions for specific services and/or frequency bands into one protocol stack, which might then be transmitted on different air interfaces (AIs). There is a trade-off between harmonization and specialization of functionality for different band services and cell types in 5G [136]. Harmonization is discussed in [148] with an emphasis on harmonization of evolved LTE and NR. Harmonization among the 5G AIs could take place at any layer (MAC, RLC or packet data convergence protocol (PDCP)), as long as the protocol stacks have the same structure at the harmonization layer and above. When the stack parameters are also the same, the layers can be aggregated, not just harmonized. Aggregation allows a single instance of the aggregated function.

Harmonization at the PDCP layer allows different AIs (LTE and NR) to be more easily coordinated, whereas harmonization at the MAC layer allows flexible scheduling of different services. To enable MAC layer harmonization, the network scheduler must be aware of the different nodes in the system, including their use cases and link conditions [148]. Each node needs to know the requirements of all the services in its harmonized stack and the resources scheduled to its different AIs so that the harmonized MAC layer can operate the MACs of different AIs and efficiently allocate the pooled resources to each service over the different AIs. Through MAC harmonization, the MAC behavior can be configured to accommodate different MAC schemes (contention-based, scheduled), different algorithms (request-to-send/clear-to-send (RTS/CTS), LBT), and different parameters (timing, resource locations, prioritization) [148]. In the context of LTE accessing the unlicensed spectrum, where the available resources are highly non-stationary, such combined dynamic cross-carrier scheduling each TTI could help fulfil URLLC requirements, although it is likely to be very challenging due to the different frame structures.

IV. LTE MAC MECHANISMS FOR URLLC

In this section we consider MAC layer mechanisms for URLLC in the licensed and unlicensed spectra. We then survey developments in the vehicular networks use case, and discuss mechanisms for LTE/Wi-Fi coexistence in the unlicensed spectrum. A summary of the MAC-layer-related URLLC techniques is given in Table V.

A. LTE MAC Mechanisms for URLLC in Licensed Spectrum

For a bearer at a UE to receive resources from an evolved node B (eNB), the bearer needs to request access, be admitted, and then have resources scheduled to it. Each of these steps has the flexibility to allow particular bearers to be prioritized, or streamlined, providing functionality that can help enable URLLC. Access class barring (ACB) provides congestion control, where low-priority bearers randomly postpone their access requests, which provides high-priority bearers with minimally contested access. The allocation and retention priority (ARP) of each bearer informs bearer admission and allows high priority bearers to usurp the channel from lower priority bearers. Once admitted, resources can be scheduled to accommodate each bearer's throughput and delay requirements, which may be through dynamic scheduling or SPS. Some URLLC are between closely located devices, such as in factory automation or between vehicles on the road. The nominal centralized process is for a device to transmit its data to the eNB and the eNB to then forward the data to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE V
SUMMARY OF MAC TECHNOLOGIES FOR URLLC

Topic	References	Features
LTE access	[149], [152], [154], [156], [157]	• Reduced RACH latency and congestion for high priority users, by retarding RA for low priority users, via (ACB, EAB) • Contention-free RA (dedicated preambles, preamble combinations) • Pre-emption (jump-in) capability for high priority bearers (ARP) • extra RACHs with many-to-one linked resources, for one-stage RA with reduced latency, at cost of reliability • eNB selection, via reinforcement learning, to increase access reliability
LTE scheduling	[158], [159], [162], [165], [166], [168]–[172]	• Definable QoS requirements (QCI) • Prioritized scheduling (QoS-aware schedulers) • SPS to remove control overheads • Pre-scheduled SPS retransmissions (pooled, multiplexed stream) • Predictive grants, to remove control overheads
D2D	[26], [57], [146], [147], [156], [173], [176]–[178], [197]	• Low-power, single-hop, user-plane transmissions, reducing latency and improving reliability • Overlay/underlay resourcing options (respectively more efficient/reliable) • Out-of-coverage access, via resource pooling, increasing reliability • With SPS for V2X applications, allowing periodic safety messages
Wi-Fi QoS	[137], [179], [190], [210]–[212]	• Contention-free polling (PCF, HCF) • Prioritized contention-based access (EDCA) • Exchange intent to Tx/Rx, reducing hidden node problem and impact of collisions (RTS/CTS) • Virtual carrier sensing to increase reliability (NAV)
DSRC	[190]–[195]	• Dedicated spectrum with defined periodic time for safety messages • Wildcard BSSID to avoid connection/association time • Distributed, (almost) contention-free protocols
LB-LBT	[176], [198], [199], [215], [219], [222]–[227], [231]	• Access to unlicensed spectrum • Multichannel diversity (Type-A multi-carrier access) • Prioritized backoff access entities (like Wi-Fi EDCA)

local recipient device. Instead, D2D communications bypass the eNB and reduce the number of hops.

1) Congestion Control - Access Class Barring: When an UE needs to send a buffer status report (BSR), for example when an UL channel buffer becomes non-empty, and the UE has no PUCCH resources available, the UE initiates a RA procedure. RA occurs on the physical random access channel (PRACH) and can be contention-based or contention-free [149]. In contention-based RA, to initiate a SR, the UE transmits a randomly chosen preamble from a set of up to 64 orthogonal preambles. If two UE select the same preamble, a collision occurs. In contention-free RA, the eNB transmits a dedicated preamble signature to a UE, enabling a collision-free SR without need for confirmation, thus improving RA latency and reliability. Dedicated preambles reduce the number of preambles remaining for contention-based RA. The preambles may also be reserved for distinct purposes, such as machine-type communications (MTC) versus human-to-human (H2H) communications [150]. Approaches to congestion control in the PRACH include: defining extra PRACH resources, dynamically defining PRACH resources, priority-based channel access, ACB, using a paging system to invite MTC devices to use the PRACH, and having group-based bearers, where one identifier allocates resources to a group, as discussed in [151].

To enable congestion control, LTE access classes (ACs) have been defined, currently numbering 0 - 15. ACs 0 - 9 are randomly assigned to all UE. AC 10 is for emergency calls and ACs 11-15 are for specific high-priority users. Admission for ACs 10 - 15 is either barred or not, whereas admission for ACs 0 - 9 can be controlled stochastically by either ACB or extended access barring (EAB). In ACB the eNB broadcasts an access probability (APr) and an AC barring time, which can be different for different ACs. When a UE attempts RA, it chooses a random number on [0, 1] and if the number is lower than the APr, the UE may transmit its RA preamble. Otherwise, the UE defers for its AC barring time and then starts over, choosing another random number. In EAB, the eNB broadcasts a barring bitmap for ACs 0-9 and ACs with their bit set in the bitmap are not allowed to initiate the RA procedure until the bit is unset. The performance of EAB for MTC is modeled in [152]. While ACB and EAB provide prioritized access for special groups, rather than for URLLC bearers, the class definitions could be altered to provide prioritization functionality to URLLC bearers more generally.

Other proposed congestion control measures include combining preambles for high-priority users to provide prioritized RA, rather than delaying access attempts of low-priority users [136], which is part of the low latency design paradigm given in the 5G METIS II vision [153]. [149] offers a reinforcement learning algorithm for MTC devices within range of multiple eNBs to efficiently select the eNB that is least congested, aiming to obtain the lowest delay. A scheme is proposed in [154], for IoT networks comprising many devices with small packets, in which multiple RBs are used for RA, and the resulting many distinguishable preambles are pre-linked many-to-one to resources. A one-stage RA is offered in which a UE sends a preamble and then transmits in the linked resource. There is no signaling, except for an ACK when the transmission is successful, so latency is very low; however, collisions are likely. A two-stage RA is also offered in which resources are specified after each SR. This reduces the collision probability, but incurs additional delay.

2) Admission Control, ARP, Pre-emption: If adding a new user to a cell pushes the demand too far, the QoS for all bearers may suffer and guaranteed QoS requirements may not be met. Admission control aims to prioritize bearers. Generally,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

a bearer is admitted if there is enough bandwidth to support the bearer, and once a bearer is admitted, its service continues, however this is not always the case.

Admission control functionality can help enable URLLC, and is a central concept in MCPTT. The idea of MCPTT is that for a high priority user to gain channel access, the user need only push a button to gain immediate access. To achieve this functionality, when an urgent message needs to be transmitted, the radio suspends its ongoing transmissions and transmits the urgent message, which is referred to as pre-emption or interruption. The decision of whether to admit a data bearer is based on its ARP. The ARP defines a priority level (1 - 15), a pre-emption capability and a pre-emption vulnerability [155]. Currently the interruption occurs at the end of the current packet [156]. In the broader URLLC context, there have also been proposals to interrupt transmissions mid-packet [157], which then require mechanisms to efficiently recover the interrupted message, such as transmitting just the untransmitted portion of the data after the interruption.

3) Scheduling: DL assignments and UL grants are scheduled by the eNB MAC layer. Allocations can be designed for different purposes, such as maximizing throughput, preserving fairness among users, and/or satisfying QoS requirements. The logic of which bearer to prioritize and under what circumstances depends on each bearer's QoS requirements, which are categorized by the QoS class identifier (QCI). The QCI specifies delay and loss targets, the priority level, and whether the bearer requires a guaranteed bit rate. The QCI classes are defined in [155] and the LTE MAC procedures are specified in [158].

DL schedulers proposed in the literature are surveyed in [159]. Schedulers can be QoS-aware, or not, and channel-aware, or not. RBs are usually allocated to UE based on per-RB metrics, which are evaluated each TTI for each bearer. The bearer with the highest metric for a RB is allocated that RB. Channel-aware schedulers have the potential to create win-win outcomes by allocating each RB to a radio bearer that is currently experiencing good channel conditions for that RB. QoS-aware schedulers deliver more win-lose outcomes, since one bearer is given priority over another. For URLLC, we are interested in QoS-aware schedulers, designed to provide guaranteed delays, which may also be channel-aware so as to provide efficiency gains and increase reliability. Such schedulers can be based on other schedulers, so we briefly describe a selection of relevant schedulers.

QoS-unaware scheduler examples include the blind equal throughput (BET), MT and proportional fair (PF) schedulers. The BET scheduler aims to provide fairness to all UE by equalizing their exponentially weighted moving average (EWMA) throughput. The MT scheduler maximizes the total throughput by giving priority to bearers with the highest expected data rate. The PF scheduler combines the BET and MT schedulers, aiming to temper maximizing total throughput with achieving fairness between UE. The corresponding priority metrics are: $m(BET)$ = 1/ (EWMA throughput), $m(MT)$ = expected data rate, and $m(PF) = m(MT) \times m(BET)$.

The earliest deadline first (EDF) and largest weighted delay first (LWDF) are QoS-aware/channel-unaware schedulers. The EDF priority metric is $m(EDF)$ = 1 / (time till deadline), where the deadline is the MAC delay threshold. So, as a packet's deadline approaches, the packet's priority rapidly increases. The LWDF priority metric is $m(LWDF) = -\log(\alpha)$ / delay threshold $\times$ HOL-packet delay, where α is an acceptable probability of exceeding the delay threshold and HOL means head-of-line. While flows are prioritized by α, LWDF does not create a scheduling imperative at the delay threshold.

QoS-aware/channel-aware schedulers are designed to maximize channel efficiency while guaranteeing delay constraints. Examples include modified LWDF (M-LWDF) [160], exponential/proportional fair (EXP/PF) [161], and the frame level scheduler (FLS) [162]. The M-LWDF combines the LWDF scheduler with the PF scheduler to give priority metric $m(M\text{–}LWDF) = m(LWDF) \times m(PF)$. However, $m(M\text{–}LWDF)$ still only increases linearly over time, with no rapid increase in priority as a packet's delay threshold approaches, so some packets fail to meet their delay requirements, which is not ideal for URLLC. The EXP/PF scheduler replaces the $m(LWDF)$ component with an exponential function of $m(LWDF)$, but still does not give particular weight to impending deadlines, so again is not ideal for URLLC.

The FLS determines the resources needed to fulfil delay-constraints over each 10 ms frame. The required resources are then allocated at each TTI using $m(MT)$, until the delay-constrained flows are all resourced, and then the remaining RBs are allocated using $m(PF)$. As such, there is the potential to optimize allocations over a longer period (10 steps), combined with the ability to dynamically adjust allocations to better utilize channel conditions, or accommodate changing demand, followed by a balancing towards fairness with any remaining resources, once constraints have been met. Noting its additional computational complexity, the FLS has potential for DL URLLC.

LTE UL scheduling is surveyed in [163] and from a machine to machine (M2M) perspective in [164]. To inform the eNB of the UL channel quality, the UE transmits sounding reference signals (SRSs). A QoS-Aware/Power-Efficient scheduler is presented in [165] where a binary integer programming problem is formulated to minimize the transmission power while meeting QoS constraints, which is then approximately solved using a greedy search algorithm to reduce complexity.

While the eNB controls the resources allocated to each UE, each UE assigns the resources across its bearers. The radio resource control (RRC) layer specifies for each logical channel the priority, prioritized bit rate (PBR), and bucket size duration (BSD) [158]. Each TTI, a variable, B_j is incremented by TTI $\times$ PBR, being the amount of data that needs to be delivered to keep up with the PBR. B_j can go negative when data is sent ahead of the PBR. Resources are allocated, in order of priority, to logical channels with $B_j > 0$, and each B_j is decremented accordingly (possibly going negative). Then, if further resources are still available, data is allocated, again in order of priority, to all logical channels with data (B_j will already be non-positive). As such, if sufficient UL resources are allocated to satisfy a high priority bearer, the UE allocation process will fulfil the high priority bearer's requirements first, as is needed for URLLC.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

4) Semi-persistent Scheduling: Standard SRs are used to request resources for every single use. Resources are then allocated each TTI with dynamic scheduling until the request is fulfilled. An alternative option is to request recurring resources, which are then allocated with one SPS grant. A SPS grant allocates periodic resources to the UE. The grant continues until it is updated, which might occur when the channel conditions change, or when the grant is terminated, either by the UE or the eNB. SPS reduces latency by removing much of the delay associated with establishing each UL grant, in the form of control signaling overhead, which has great potential for URLLC. SPS also reduces the collisions in the PRACH, since the number of SRs is reduced, thus reducing SR latency and reliability for single-use resources.

The SPS periodicity is preconfigured via the RRC signal, and then the allocated RBs and MCS are conveyed through the PDCCH [166]. SPS uses persistent scheduling for initial transmissions and dynamic scheduling for retransmissions, in the event that the initial transmission fails and no ACK is returned. The distinction between persistent scheduling and SPS is made in [167]. Persistent scheduling occurs at protocol stack layer 3 and uses a fixed MCS, which reduces control needs, but does not adapt to the changing channel quality.

The reduction in latency provided by SPS is demonstrated in a number of papers. [168] demonstrates that fast UL grants, using SPS, can decrease UL transmission latency from 12.5 ms to 7.5 ms for LTE Release 14 settings. [169] explores SPS in a factory setting and demonstrates that SPS reduces latency to less than half that of dynamic scheduling. Instant UL access (IUA) is also described in [169] as a low latency mechanism for sporadic, yet delay sensitive, traffic. In current specifications, the UE is expected to send a MAC PDU in response to an allocated dynamic UL grant or configured SPS grant, even if no data is available for transmission [64]. IUA is very similar to SPS, in that periodic resources are reserved; however, the resources are not expected to be always used. Hence, the lower latency comes at the cost of potentially high spectral inefficiency.

Several ideas related to SPS and reducing control overheads have potential to further reduce latency, including:

- pre-scheduled SPS retransmission resources, that can be pooled between several users and reallocated, if not needed [170];
- developing this point, to avoid over-resourcing, we can propose to conduct persistent scheduling coupled with traffic multiplexing in the upper layer, where flows are disassembled into multiple simultaneous sub-flows, or parallel bearers, that are then transmitted/retransmitted over a single persistent resource stream that is adaptively resourced so as to maintain a consistent data rate for the overall traffic flow; and
- predictive resource grants, e.g., in a factory setting where if machine B always transmits a set time after machine A transmits, an UL grant is offered to both machines A and B when machine A makes a SR [171], or when an application layer triggers a service request, (at the non-access stratum (NAS) layer), bearers are established without there being data in the UE buffer, based on learnt data activity relationships between service contexts [172].

5) Device-to-device: D2D communications have the ability to reduce end-to-end latency, by creating direct connections between closely located source and destination nodes. D2D can also improve reliability thanks to the insignificant path loss in short range communications, as illustrated in Fig. 10. By removing the eNB from the data route (and possibly the core network when UE are in different cells), D2D communications can also produce higher data rates and better resource efficiency [173]. D2D emerged through the development of LTE public safety communications functionality. Release 12 D2D features include support for broadcast [174] and Release 13 D2D ProSe features include supporting priority control, out-of-coverage discovery, and UE-to-network relay [156].

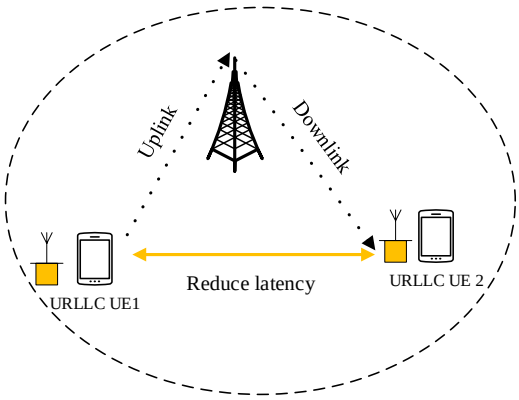


Fig. 10. D2D illustration.

D2D communications can operate in either scheduled or autonomous mode [175]. In the scheduled mode, D2D resources are scheduled by the eNB, which requires the UE to be in-coverage. Resources are allocated with dynamic scheduling, except in the case of V2X, where SPS may be used instead. In the autonomous mode, a UE selects resources from a resource pool, which is indicated in the system information block (SIB) when in-coverage, and can be preconfigured for use when out-of-coverage, creating immunity to network failure. The pooled resources follow binary on/off subframe transmission patterns of length 5-8 subframes, which then repeat [176]. This also provides on/off patterns with periodicities of 1-4. It is argued in [26] that, as well as the D2D latency advantage stemming from single hops, the out-of-coverage functionality is essential for reliable vehicular communications. Prior to a UE transmitting data on a D2D channel, or sidelink, the UE informs the receiving UE about the transmission(s) by transmitting sidelink control information (SCI). There is no feedback for the SCI, so it is transmitted both with control messages and data transmissions to increase reliability [156].

Since D2D links, or sidelinks, can be considerably shorter than the distance to the eNB, there is the potential to reduce the transmission power. [177] surveys D2D interference management. Inband D2D resource allocations can be in the underlay or overlay mode. In the underlay Mode, D2D pairs and cellular UE share the same spectrum resources, and interference is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

limited by controlling the transmission power of both D2D and cellular links. In the overlay Mode, dedicated spectrum resources are allocated for D2D communications. A dynamic resource allocation scheme that supports the D2D underlay mode is offered in [178], based on the channel gains between the various links.

A novel architecture is proposed in [146] that combines C-RAN and D2D to handle the associated fronthaul delay of C-RAN. The authors believe this combined architecture can be deployed in licensed and unlicensed bands with almost "zero delay", helping to solve the delay issue and fulfil most of the targets specified for 5G networks. A similar link-reduction idea is caching, in which, if multiple users require the same information and it is stored locally (at another UE, BS, or macro base station), the information can be retrieved with fewer hops [57].

The effects of heterogeneous user and device mobility are studied in [147]. D2D connections and drone-assisted links are highly utilized and improve the availability and reliability of URLLC data acquisition at low and moderate device speeds. Simulations show that improvements of up to 40 percent in reliability, compared to a cellular-only baseline, could be obtained. However, when the speed of movement increases, the improvements provided by D2D links and drone small cells decreases.

B. URLLC MAC Mechanisms for Incumbent Technology in Unlicensed Spectrum

1) Wi-Fi: Wi-Fi, or IEEE 802.11 [137], is the incumbent technology in the unlicensed spectrum. The legacy IEEE 802.11 channel access mechanism, known as carrier-sense, multiple access with collision avoidance (CSMA/CA), is a distributed coordination function (DCF). CSMA/CA is slotted, with the channel's busy/idle status determined by carrier sensing.

The carrier sensing synchronizes the MAC slots of the access processes from all the independent Wi-Fi stations, in a distributed manner. When the channel is sensed busy, the MAC slot is deemed 'busy', and the slot continues until the channel is sensed idle during a continuous DCF interframe space (DIFS). At this time, all the Wi-Fi stations should be synchronized. Subsequently, each 'slotTime' during which the channel is sensed idle counts as an 'idle' MAC slot, again resulting in all the Wi-Fi stations being synchronized. For the IEEE 802.11n and IEEE 802.11ac protocols, the DIFS is 34 μs and the slotTime is 9 μs.

When a Wi-Fi station is in an idle state and a packet arrives, the station first senses the channel for a DIFS. If the channel is sensed idle, the station transmits. Otherwise, the station enters backoff stage-0. In backoff stage-0, an integer counter is drawn for a uniform distribution over [0, CW_0], where CW_0 is the contention window (CW) size for backoff stage-0. After each MAC slot, either busy or idle, the counter is decremented, and once the counter reaches zero, the station transmits.

If the first transmission attempt fails, i.e., is not acknowledged, the station enters backoff stage-1; if then the second transmission attempt fails, the station enters backoff stage-2; etc. Each backoff stage is the same as backoff stage-0, except the 'CW size plus one' doubles, up to a limit. The CW for backoff stage-i, CW_i, is given by $CW_i = (CW_0 + 1) \times 2^{\min(m,i)} - 1$, for $0 \leq i \leq s$, where m is the doubling limit and s is the retry limit. If a transmission in backoff stage-s fails, the packet is dropped by the MAC layer and reported to higher layers.

After either a successful transmission or $s + 1$ failed transmissions, the process returns to backoff stage-0, regardless of whether the station has another packet to transmit. When the counter next reaches zero, the station either transmits, if it has a ready packet, or otherwise enters the idle state.

As CSMA/CA is a contention-based access mechanism, collisions are inevitable. In a congested channel, the MAC delay can reach 100's of milliseconds. To alleviate the collision implications on the access delay and reliability, RTS/CTS is an optional two-way handshake mechanism [179]. Specifically, stations inform nearby nodes about their incoming transmissions and the nodes set their network allocation vectors (NAVs) to the transmission duration as a form of virtual carrier sensing. If all transmitting devices use RTS/CTS, collisions only occur on the RTS frame, so collision durations can be vastly reduced [180], especially for long aggregated packets. Data frame headers also include the transmission duration, so are similarly used to set NAVs and improve reliability. Wi-Fi amendments introduced MIMO in 802.11n, multi-user MIMO in 802.11ac, and 60 GHz millimeter-wave operation in 802.11ad.

IEEE 802.11ad supports 6.75 Gbps over distances of up to 10 m [181]. Beamforming is used to reach specific destinations and overcome attenuation. 802.11ad has a superframe comprising a beacon transmission interval, data transfer interval (DTI), and optional association beamforming training (A-BFT) or announcement transmission intervals (ATI). The DTI can have scheduled service periods (SP) and contention-based access periods (CBAPs). Spatial sharing is possible, enabling simultaneous SPs. The reliance on directional transmissions creates the deafness problem [182], which is similar to the hidden node problem. A centralized directional CSMA/CA protocol, which extends directional MAC (D-MAC) [183], is proposed to alleviate the problem [182]. Target RTS frames are directed to a piconet coordinator (PNC) and target CTS frames, containing sender/receiver locations and transmission duration, are returned omnidirectionally. The sender and receiver then steer their antennas at each other for high speed transmission while other devices set their NAVs. An alternative dual-band CSMA/CA approach is proposed in [184] in which omnidirectional control messages are exchanged on 5 GHz Wi-Fi frequencies, being highly reliable over short distances, and data frames are exchanged on the high speed 60 GHz frequencies. The deafness problem is mitigated and high speed resources are spent on data transmission, not control messages. The dual-band approach of [184] achieves frame delays under 0.01 ms, for 15 KB frames and throughput over 2.5 Gbps for 32 stations within a 23 m radius of the AP.

2) Wi-Fi Quality of Service Mechanisms: To address the lack of QoS guarantees, the Wi-Fi 802.11 standard [137] includes two channel access mechanisms that an AP can use to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

centrally control Wi-Fi traffic, the point coordination function (PCF) and hybrid coordination function (HCF). Both use polling with priority defer periods, of duration PCF interframe space (PIFS), to create contention-free periods (CFPs), which alternate with contention periods (CPs). The PCF polls STAs in turn, working through a polling list, whereas the HCF makes QoS-aware polling decisions based on QoS STAs' parameters, and can also instigate polling during the CPs.

The HCF additionally defines four priority ACs for contention-based channel access during the CPs, as part of the enhanced distributed channel access (EDCA) mechanism. Higher priority ACs have shorter backoff processes, created by shorter CWs and shorter arbitration interframe spaces (AIFSs). The AIFS replaces the DIFS after a frame exchange and has timing such that PIFS < DIFS = AIFS[highest priority] < AIFS[lowest priority]. The two highest ACs also have a specified maximum transmission opportunity (TXOP), during which multiple packets may be sent. A single station (STA) can have a separate backoff entity for each of the four ACs and when multiple backoff entities from the one STA are due to transmit simultaneously, the highest AC backoff entity transmits and the other backoff entities incur a virtual collision. While these mechanisms do not provide QoS guarantees, they can provide differentiated QoS for different ACs and greatly improve the latency and reliability for high priority flows. From the perspective of LTE accessing the unlicensed band, long CFPs pose an access problem.

C. Study Case: Intelligent Transportation System

ITS will rely on V2V and V2I communications. Vehicular communications links require low latency for safety and can be short lived due to high mobility. In 1999, the federal communications commission (FCC) allocated 75 MHz of bandwidth (from 5.850 GHz to 5.925 GHz) to DSRC for the vehicular environment [185]. DSRC operate under the IEEE 802.11p/1609 wireless access in vehicular environment (WAVE) protocols.

Instead of using DSRC for vehicular communications, another option is to rely on the LTE protocols, which offer both long-range links and D2D sidelinks. A further option is to combine the two. [185] surveys heterogeneous vehicular network (HetVNET), focusing on the MAC layer. HetVNETs integrate DSRC and cellular networks to enable the ITS. There are a number of different options for HetVNET architecture: DSRC for V2V and V2I, Cellular D2D for V2V, Cellular broadcast/multicast for V2I, and Cellular unicast for V2I. [185] finds that V2I communications are better served by LTE because of the wide coverage, high capacity, robust mobility management, and centralized architecture; whereas V2V communications are better served by DSRC because of easy deployment, low costs, the ad-hoc mode and low WAVE short message overheads. As such, an integrated and collaborative HetVNET that takes advantage of both protocols is seen as essential for a functional ITS. The internetworking of DSRC and cellular technologies, including hybrid architecture and handover options, is surveyed in [186].

1) Dedicated Short-range Communications: The IEEE 802.11p standard defines the DSRC channel access rules and time divides the channel into alternating CCH intervals (CCHIs) and SCH intervals (SCHIs), each nominally 50 ms and starting with a 4 ms guard interval. Consecutive CCHI and SCHI pairs form a periodic 100 ms synchronization interval (SI). The IEEE 1609.4 protocol [187] extends IEEE 802.11p to multiple 10 MHz channels, one CCH and six SCHs. The basic idea is that all devices participate in the CCH during the CCHI, so that devices can announce themselves (position, speed, acceleration, and direction), find other devices, and receive WAVE service advertisements (WSAs). The WSAs are transmitted by roadside units (RSUs), announcing the services offered on the SCHs, and are covered by the IEEE 1609.3 protocol. The DSRC standards are overviewed in [188], including the IEEE 802.11p amendment for WAVE, the IEEE 1609.3 standard for network services and the IEEE 1609.4 standard for multi-channel operation. Potential VANET (Vehicular Ad-hoc Network) MAC protocols and applications are overviewed in [189].

The usual Wi-Fi architecture is to form a basic service set (BSS), wherein an AP, which is directly connected to the network, communicates with multiple stations. In DSRC, vehicular safety messages are usually broadcast instead. Wi-Fi broadcast is achieved by setting the BSS identifier (BSSID) to the wildcard BSSID (0xFFFFFF), which addresses packets to all nodes also using the wildcard BSSID and does not require synchronization, authentication, or association before transmissions commence. Broadcasting reduces the number of links, so reduces congestion and has the potential to reduce latency; however, since the transmissions are addressed one-to-many, they are not directly acknowledged, so the reliability is reduced.

When large numbers of vehicles are within range, DSRC suffers serious channel congestion [185]. This is a concern for safety channels and finding methods to mitigate the problem is an area of research. The IEEE 802.11p protocol includes EDCA to provide higher priority for safety messages and control messages. [190] modeled EDCA in a V2V environment and found that vehicular latency requirements can be met for the highest priority access class (AC), although reliability requirements are not met. Interference and collisions can be reduced by adaptively controlling the transmission range and rate [191], relay retransmission probabilities [192], and CW and transmission power [193]. The common theme being that as the channel occupancy increases, the communication range is decreased to maintain an acceptable channel capacity. Also, the hidden node problem is a particular concern, due to the linear nature of highways, which can be mitigated by setting the communication range shorter than the carrier-sensing range.

Other methods for DSRC have been explored to improve the reliability of CCH safety messages. In [194], a decentralized cluster head approach, meaning one vehicle from a vehicle cluster coordinates the cluster, produces contention-free allocations and good reliability, with the safety message exchange failure rate kept to 3×10^{-5} in simulations, regardless of the non-safety message load. A self-organising distributed scheme,

dedicated multi-channel MAC (DMMAC), is proposed in [195] that is similar to reliable reservation ALOHA (RR-ALOHA) [196]. A vehicle reserves a slot it deduces available from a slot-status table, and, as well as transmitting safety data during its reserved slot, transmits its understanding of the status of all slots, so that other vehicles can update their slot-status table. In a simulation with increasing traffic density, DMMAC demonstrated good reliability, achieving periods with near 100% packet delivery, interlaced with brief dips down to 80% packet delivery as the network topology changed and the scheme reestablished, whereas under the WAVE MAC, packet delivery fell from 90% to 20%.

2) LTE V2X: In [197], groups of vehicles, or platoons, use underlay D2D SPS to communicate location information within their platoon, forming a super frame, while the platoon leader communicates with the core network, to exchange information externally. The platoon leader also coordinates the pooled resources for the platoon, cycling transmission opportunities through the platoon members faster than every 100 ms, which satisfies V2V HRLLC use cases. Short distances and low transmission power allow spatial reuse of resources. SPS functionality for V2X sidelinks is available in 3GPP Release 14 [158], with peroidicities including 20 ms, 50 ms and 100 ms.

D. Coexistence in Unlicensed Spectrum

The unlicensed spectrum is a large source of additional bandwidth with the potential to greatly enhance licensed spectrum communications. However, by its nature, the unlicensed spectrum does not provide exclusive access to any particular operator and, as such, the availability of unlicensed resources cannot be guaranteed at a chosen time. Current research efforts have focused on creating mechanisms that enable fair coexistence between different technologies competing in the unlicensed spectrum, where fair coexistence has predominantly been assessed in terms of channel time and throughput. As such, there is little current research addressing URLLC in the unlicensed spectrum. If the unlicensed spectrum is to support URLLC, more focus is needed on mechanisms for latency reduction and successful delivery.

1) LTE Access to the Unlicensed Spectrum: The LTE protocol has been developed to access to the unlicensed spectrum via the LTE air interface, allowing LTE traffic to be coordinated under a single framework with the rich set of LTE features. Access is via carrier aggregation with the unlicensed channel treated as a secondary cell (Scell) and the primary cell (Pcell) being in the licensed spectrum. Due to the contention-based nature of the unlicensed spectrum, most of the LTE control signaling is expected to be sent on the Pcell, although control signals can be sent on the Scell [140]. LTE access mechanisms to the unlicensed spectrum can be broadly split into being frame based or load based.

In frame-based access, channel access is made or attempted periodically at chosen times that are convenient, such as at the start of an LTE frame or subframe. Frame-based access may be a deterministic, creating an on-off duty cycle, so as to obtain access for a particular proportion of time in a particular pattern; or it may be probabilistic, with the outcome of periodic access attempts dependent on sensing the channel and finding it not in use. The duty-cycle approach is referred to as LTE-unlicensed (LTE-U) and, when the duty cycle is adaptively adjusted based on channel activity, is also known as carrier sense adaptive transmission (CSAT). The process of sensing the channel prior to transmitting is known as a clear channel assessment (CCA) and access mechanisms that apply a CCA are referred to as Listen-before-Talk (LBT) procedures. In Japan and Europe it is mandatory to use LBT in unlicensed bands. Proposed frame periods have been relatively long, such as 50 ms, comparable to a 10 ms LTE frame, or as short as 1 ms, so as to align with LTE subframes. The frame period can also be slowly adapted based on the channel activity to achieve an objective. Once access is obtained, a transmission may continue up to a maximum time, which normally is shorter than the frame period, and then the channel is released.

Load-based channel access mechanisms are generally LBT schemes. Load-based LBT (LB-LBT) procedures most commonly include a CW from which a random number is selected that defines how long the device needs to wait before transmitting. The wait is in slots, which are very short when the channel is idle, or the duration of the current transmission plus a defer period, when the channel is busy. Hence, the greater the channel activity or load is, the longer the wait between LB-LBT transmissions. In some schemes the CW is fixed, in some the CW is automatically adjusted based on transmission outcomes, and in some it is dynamically adapted to optimize an objective.

An alternative approach to cellular networks utilizing the unlicensed spectrum, when devices have both LTE and Wi-Fi air interfaces, has been to offload some of the traffic to the unlicensed spectrum through Wi-Fi APs, using LTE-WLAN aggregation (LWA). Besides requiring dual air interfaces, the implementation of LWA requires additional infrastructure, such as Wi-Fi hotspots; and then the offloaded traffic is at the mercy of Wi-Fi protocols.

2) LTE/Wi-Fi Coexistence Standards: The 3GPP standardized LB-LBT for licensed assisted access (LAA) to the unlicensed spectrum in Release 14 [176]. Concurrently, the ETSI standardized a similar LB-LBT access scheme, and has also retained a FB-LBT access scheme [198].

The two LB-LBT versions are very similar. Both are backoff based with four access priority classes, and are similar to the Wi-Fi EDCA mechanism. Higher priority classes have shorter minimum CWs, fewer backoff stages and shorter defer periods, so that access occurs more frequently. To counter the advantage, a shorter channel occupancy time (COT) is allowed each access.

More specifically, the different defer periods are equal to a DIFS +/- multiples of the idle-slot time, which means that all the nodes having sensed the channel idle for their respective defer period have synchronized MAC slot transitions. A node with a longer defer period may have to wait for nodes with shorter defer periods to transmit many times before the channel becomes idle long enough for it to recommence its backoff process. This is the underlying mechanism that gives prioritized access to specific ACs. In DL, the highest

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

priority class defer period equals a PIFS, as used by a Wi-Fi point/hybrid coordinator. UE can transmit within the COT obtained by an eNB, separated by short breaks, which is one mechanism for UL transmission. UE can also obtain access via their own backoff processes. There are again four access priority classes, which are similar to the DL set, the main difference being that an additional slot is added to the defer periods of the highest two priority classes.

The COTs are shorter than the standard 10 ms LTE frame duration for Type 1 and Type 2 frame structures. Due to the unpredictable timing of contention-based access, a new LTE frame structure, Type 3 frame structure, has been defined for LAA that allows frames of different lengths and allows for transmissions to end mid subframe [199]. The 4-bit 'Subframe configuration for LAA' field indicates the number of unoccupied OFDM symbols in the current or next subframe.

The main differences between the 3GPP and ETSI LB-LBT standards are that: the maximum COTs of the non-highest-priority ACs are different; the ETSI CW requires just one ACK to be reset to its minimum value, whereas 3GPP requires 20% percent of the HARQ responses to be ACKs; and the multi-carrier access options differ.

3GPP has Type A and Type B multi-carrier access. In Type A, a separate backoff processes is maintained for each carrier. This appears to have potential for reducing the periods without access. However, it is implied that once one backoff process gains access and a transmission starts, the other backoff processes are paused, which reduces the potential gain. The ETSI equivalent, Option 1, allows the backoff processes to be independent, if the devices are capable of maintaining independent transmissions on separate carriers.

In Type B, a single backoff process is maintained on one carrier and used for all carriers. Just prior to transmission, all carriers in the multi-carrier set are sensed to assess channel activity. Transmissions may be made on all carriers that are found to be idle. The channel on which the backoff process is performed is either selected randomly after each transmission or selected arbitrarily, but no more often than every second. As such, although the Type B multi-carrier option potentially increases capacity, it does not reduce the periods without access to the unlicensed spectrum. The ETSI equivalent, Option 2, is like 3GPP Type B.

3) $\mathrm{Multefire}^{\mathrm{TM}}$*:* Unlike the 3GPP and ETSI standards, which are anchored in the licensed spectrum, Qualcomm has developed $\mathrm{Multefire}^{\mathrm{TM}}$ as a standalone technology in the unlicensed spectrum [200]. $\mathrm{Multefire}^{\mathrm{TM}}$ introduces the idea of a Neutral Host that can support other mobile network operators, acting as a central scheduler that balances channel loads. It can also coexist with other users and operators, either dynamically selecting and aggregating clear channels, or otherwise using LBT to fairly share occupied channels.

$\mathrm{Multefire}^{\mathrm{TM}}$ relies on frequent transmissions of an enhanced discovery reference signal (DRS). The DRS occupies 1 ms, so only requires $T_{drs} = 25$ μs, rather than a full CCA backoff process, to access the channel. The DRS block includes synchronization signals, reference signals, CSI, and system information broadcast (SIB). The aim is to opportunistically try to send a DRS every 10 ms and periodically, either 40, 80 or 160 μs, persistently try to send a DRS throughout a window [201]. The UL waveform is block interleaved FDMA (B-IFDMA), which divides a 10 MHz (or 20 MHz) channel into 5 (or 10) 2 MHz wide interlaces, and then each interlace comprises 10 physical RBs. The interlaces are used for PUCCH, PUSCH, PRACH and SRS. A short PUCCH (sPUCCH) and an extended PUCCH (ePUCCH) are introduced. The sPUCCH occurs in the last four symbols of a DL subframe before an UL subframe to allow UL control information to be sent before an UL transmission burst or to carry small-payload UL control information, such as ACK/NACK. The ePUCCHs carry HARQ, CSI and SR UL control information and are dynamically scheduled by the eNB, meaning that they may be interlaced with dynamically scheduled PUSCH resources. Rather than DL HARQ being fed back in frame $n+4$, it is fed back in a bit map at the earliest UL TXOP with control information resources, after allowing four subframes for processing. PRACHs are similarly interlaced, with one or more interlaces being dynamically configured as available by the eNB.

V. Evaluation of URLLC Enabling Technologies

To achieve URLLC, all steps through the communications process can be streamlined or sped up to reduce latency. At the PHY layer, this can mean using a frame structure with shorter symbols and shorter TTI and using a waveform that allows quicker processing, both for coding at the transmitter and decoding at the receiver. At the MAC layer, this can mean removing control signaling delays and scheduling bearers in the best order. Improvements on either aspect will aid URLLC. Reliability can be improved by measuring the channel quality, so that error-rate trade-offs can be selected; by diversification, so that stochastic variations can be reduced and SINR increased, and by including other redundancy, such as increasing coding and frequency resources.

In this section we consider simulations and a lab trial that combine multiple PHY-layer URLLC components to create feasible URLLC systems. We then consider MAC-layer modeling and simulations, with a focus on utilizing the unlicensed spectrum within a URLLC solution. The priority classes within the 3GPP LAA protocol are very similar to the Wi-Fi EDCA, so we survey EDCA modeling efforts. Then we survey modeling and simulations of LTE access to the unlicensed specturm, i.e., LTE/Wi-Fi coexistence mechanisms.

A. PHY-layer Evaluations

Evaluation metrics and evaluation methods for URLLC are defined in detail in [32]. Multiple simulation studies are presented in the 3GPP technical report [64], evaluating latency reduction techniques in terms of link-level, system-level and design aspects. The study flags reduced SPS periodicity, fast UL grants, and reducing TTI and processing times as beneficial to latency reduction.

It is improper to adopt existing LTE/LTE-A evaluation methodology in the URLLC case directly, since there are several challenges emerging with several critical requirements

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

in the ultra-low BLER condition, such as, more precise propagation models, multi-targets with more parameters, and simulation accuracy/efficiency. Thus, novel methodologies should be considered in the PHY evaluation of URLLC. In [202], the authors propose a cloud-based two-level network simulation framework, which acquires scalability, flexible resource calling, and self-management, to improve the evaluation efficiency in various simulation scenarios. Paper [203] investigates the candidate test environments for 5G evaluation, and the authors point out the indoor isolated environment is an appropriate modeling for the URLLC usage scenario. In addition, an initial system-level simulation with the multiple-user coherent joint transmission is carried out to evaluate the 5G performance in the indoor isolated environment, showing that the significant gains in SE and reliability come from diversity, cooperation, and interference suppression.

Most of the papers related to URLLC components discuss, analyze, and/or characterize the particular URLLC component being explored, but do not simulate or test a full URLLC system. Table VI summarizes recent evaluations of individual PHY-layer URLLC techniques. Nonetheless, there have been sufficient simulations and lab trials to demonstrate that URLLC is feasible, given sufficient resources. The issue then is how to achieve the URLLC requirements for the lowest cost, in terms of bandwidth, infrastructure, energy and lost performance for non-URLLC users.

The SINR distribution is simulated in [204] for a selection of diversity techniques where multiple BSs are in a network. As well as implementing MIMO, which creates microscopic spatial diversity, the BSs can cooperate by either simultaneously transmitting the same signal, to create macroscopic spatial diversity, or perform interference cancellation, and/or implement frequency reuse, where frequency resources are kept segregated between different sectors, to reduce cochannel interference. Combining microscopic spatial diversity, macroscopic spatial diversity, and frequency reuse satisfied the 99.999% reliability target most convincingly. However, for practical implementation considerations, a 4x4 MIMO scheme with second-order macroscopic diversity was advocated the most feasible configuration.

Short TTI, together with multiple antennas are able to transmit small packets (100 bits) with high reliability, using low-rate MCS (1/2 rate coding, QPSK) [27]. A diversity order of 16 reduces the tail effects of Rayleigh fading from 90 dB to 9 dB, so that in a factory setting (having short distances and multiple base-station antennas mounted to the ceiling), where coherent combining at the receiver producing further processing gains, packet error rates of 10^{-9} are feasible. The issue then becomes how much bandwidth and energy to spend to achieve such reliability. Related simulation results are presented in [205], for which a BS is centered within a 100 m x 100 m factory environment and UE dropped randomly, while allowing a minimum 5 m link distance. The design included: reduced TTI and shorter OFDM symbol durations; use of convolutional codes instead of turbo codes; physical channels that enable early channel estimation; and high diversity levels. In addition to background noise and attenuation, different interference levels were simulated to represent other nearby factories. Trade-offs between latency, MIMO configuration, and the required bandwidth, while providing 99.999% reliability were given. Using 2x8 MIMO to transmit 1000-bit packets, the cost of the interference power doubling is approximately a 50% increase in required bandwidth.

A lab trial was reported on in [206], which achieved a 1.5 ms HARQ RTT in DL, including the processing time for the next DL transmission. The trial used 2x2 DL MIMO and a TDD frame structure that had: 30 kHz subcarrier spacing; 0.25 ms subframes; and expected HARQ feedback in the latter portion of alternate subframes. For high SNR (25-26 dB), 328-bit (41-byte) packets that were transmitted with 16 QAM had 100.00% decoding rates, which translates into at least 99.995% reliability within a 1.5 ms latency, given high SNR. The trial was conducted between one BS and one UE in an anechoic chamber, and did not account for control overheads, such as scheduling or CSI estimates. The achieved RTT is around 5 times shorter compared with what is supported by LTE-Advanced standard and shows the feasibility of ultra-low latency in 5G.

However, published evaluation results are operated under various ideal or non-ideal assumptions, and they are generated from different platforms with or without calibrations. Thus, it is hard to make an appropriate comparison between the aforementioned PHY techniques just from the publications. Here, we carry out a Monte-Carlo simulation to reveal the relationship between reliability and latency for several PHY techniques, when experiencing different target SINRs.

Simulations are performed to demonstrate the reliability of different PHY techniques under the given transmission latency and SINRs. It is assumed that a 32-byte packet is appropriately encoded to be transmitted in a block whose size is 20 MHz $\times$ 1 ms. For simplify, we assume the signals experience a Rayleigh fading channel. The transmission latency is defined as the total transmission time, equal to the sum of UL and DL transmission symbol durations. The target received SINR is defined as the long-run average received SINR over a large number of transmissions.

In each iteration of the simulation, the target received SINR and transmission latency are given as constraints, and a realization of the Rayleigh channel is generated independently. The selected PHY techniques are implemented in turn and assessed on reliability, that is, the error-free rate of the 32-byte packet transmissions. In the "baseline" scenario, one antenna is equipped at each of the source and destination UE, and two antennas are equipped at BS. In the "space diversity" scenario, two antennas are equipped at the source and destination UE, respectively, while four antennas are equipped at BS. For the "frequency hopping" scheme, each transmission hops between five independent 20 MHz channels. In "accurate spectrum sensing", the equivalent received SINR can be enhanced by a factor of five. Adopting "in-band full-duplex", the transmission latency in both UL and DL is 1 ms. We assume that, transmission beamforming is used with CSI acquired at the transmitter, resulting in a power gain. Otherwise, equal power transmission is used when no CSI is acquired at the transmitter.

As shown in Fig. 11, diversity-based techniques, such as space diversity and frequency hopping, obtain noticeable gains

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE VI
EVALUATIONS OF PHY-LAYER URLLC TECHNIQUES

Technique	Ref	Contributors	Method	Main points
Frame structure	[65]	Ericsson	link-level system-level	A frame structure which revises LTE by shortening the TTI by a factor of five, from 1 ms to 0.2 ms. A coherent matched filter is good for the SR detector. Support a BLER of 10^{-9}.
	[66]	Huawei	system-level	A symbol-wise frame with reused numerology, low CP overhead and scattered pilot outperforms self-contained frame structure, especially at high Doppler scenarios.
	[68]	Qualcomm	link-level system-level	A 1-symbol based TTI has ultra-low latency operation and high system capacity.
	[206]	Huawei DoCoMo	lab trial	Achieved 1.5 ms HARQ RTT in DL with a 0.25-ms TDD frame structure.
	[205]	Ericsson	link-level system-level	Reduced TTI and shorter OFDM symbol durations, while providing 99.999% reliability.
	[64]	3GPP	link-level system-level	Reduced SPS periodicity, fast UL grants, and reducing TTI and processing times are beneficial to latency reduction.
Waveform design	[80]	Nokia	link-level	Time-domain implementation is effective in the case of single or few narrow subbands. For high number of subbands, or wide subbands, the FC-F-OFDM scheme is clearly more effective in terms of the multiplication rate.
Frequency/time/ space diversity	[86]	Ericsson	system-level	The diversity affects the system capacity markedly both in noise-limited and interference-limited scenarios
	[204]	Nokia	system-level	For a 10^{-5} desired SINR outage, a 4x4 MIMO scheme with second order macroscopic diversity is considered as the most feasible configuration when taking practical implementation considerations into account.
	[205]	Ericsson	link-level system-level	High diversity levels such as 2x8 MIMO to transmit 1000-bit packets are recommended.
	[206]	Huawei DoCoMo	lab trial	For high SNR, packets that were transmitted with 2x2 DL MIMO and 16 QAM have 100.00% decoding rates, which translates into at least 99.995% reliability within a 1.5-ms latency.
MCS	[205]	Ericsson	link-level system-level	Use of convolutional codes instead of turbo codes for user data, since they are faster to decode and do not introduce an error floor, while still achieving almost the same performance for the anticipated short message lengths.
Frequency hopping	[100]	ZTE	link-level	A sequence based sPUCCH containing only 2 SC-FDMA symbols can improve reliability by utilizing symbol-level frequency hopping.
Grant-free NOMA	[125]	Huawei	lab trial	Grant-free sparse code multiple access can significantly reduce the transmission latency in heavy overloading scenarios.
Multi-connectivity	[147]	Ericsson	system-level	Improvements of up to 40 percent in link availability and reliability with the use of proximate connections on top of the cellular-only baseline are verified.
	[144]	Nokia	system-level	A multi-connectivity concept for a cloud radio access network can considerably reduce mobility related link failures without a loss in throughput of cell-edge users.

due to frequency/space diversity. Otherwise, the in-band full-duplex technique slightly outperforms the baseline because of the lower data rate and power gain. Also, accurate spectrum sensing technique brings marked improvement from the power gain. Space diversity with accurate CSI achieves the highest gain in average received SINR, achieving approximately 25 dB more than the baseline for 99.999% reliability. Considering the spectrum and energy efficiency, frequency hopping may be the most efficient way to increase reliability under given throughput and latency constraints.

The reliability of different techniques, under a -5 dB target SINR constraint and with different transmission time constraints, is illustrated in Fig. 12. For this low average received SINR region, i.e., -5 dB, techniques enhanced by SINR gains (space diversity and frequency hopping) can achieve a reliability of 99.999% within a 1 ms latency constraint, whereas the other PHY techniques are orders of magnitude less reliable. It is noteworthy that under space diversity with accurate CSI, the reliability for a 0.1 ms latency constraint still exceeds 99.999%. Furthermore, Fig. 12 shows that the transmission latency can be reduced to below 0.4 ms with well-designed diversity-based techniques, even without precise CSI. This allows for propagation and processing delay budgets of more than 0.6 ms, while adhering to a total latency constraint of 1 ms.

Thus, we can see that, when the transmission power is increasing, reliability constantly increases under certain transmission latency and bandwidth constraints. However, to avoid severe interference, the transmission power should be limited. Therefore, diversity-based techniques have great potential to be deployed in licensed and unlicensed bands harmoniously.

B. MAC-layer Evaluations

1) Contention-based Access Markov-Chain Modeling: Contention-based access to the unlicensed spectrum has been frequently modeled using Markov chains. The seminal paper, [180], modeled the Wi-Fi CSMA/CA protocol for N Wi-Fi STAs with saturated traffic. An embedded Markov chain models the channel access protocol for a single STA and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

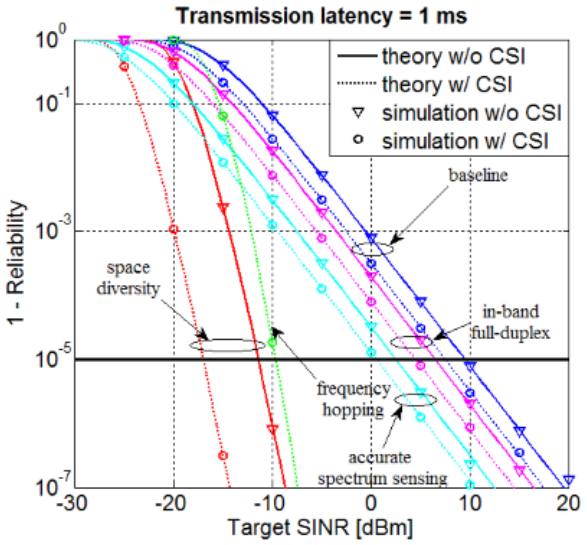


Fig. 11. Reliability vs. target SINR under a 1 ms transmission latency constraint.

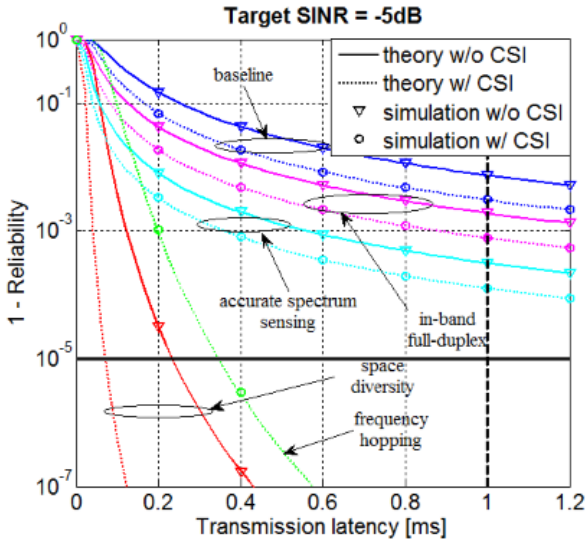


Fig. 12. Reliability vs. transmission latency under given target SINR.

a network interaction model accounts for the interactions between the STAs.

The states of the Markov chain represent MAC slots, which have different durations. The Markov chain inputs the long-term collision probability as seen by the STA, p, and outputs the STA's transmission probability, τ, which is extracted from the Markov chain's steady-state solution. The network-interaction model conversely calculates p from the τ of all the other STAs in the network, under the weak assumption that the processes are independent. For saturated traffic, all the transmissions probabilities are equal by symmetry. For the N-saturated STA case, $p = 1-(1-\tau)^{N-1}$. The single-node and network-interaction models are alternately solved to obtain a simultaneous solution.

The saturated traffic case provides performance limits under heavy congestion. A model for finite load and finite input buffer is offered in [207], where each STA has the same finite load, so that again, one Markov chain represents each STA, by symmetry. The model is extended in [208] to heterogeneous finite loads, along with transmission errors and the capture effect in a Rayleigh fading channel.

LB-LBT/Wi-Fi coexistence in the unlicensed spectrum has been modeled using two Markov chains, one that represents each Wi-Fi node and another that represents each LTE node. The network interaction model then becomes

$$p_L = 1-(1-\tau_W)^{N_W}(1-\tau_L)^{N_L-1}, \quad (4)$$

$$p_W = 1-(1-\tau_L)^{N_W-1}(1-\tau_L)^{N_L}, \quad (5)$$

where subscripts L and W respectively denote LTE and Wi-Fi nodes.

2) EDCA Modeling and Simulations: The EDCA mechanism has been simulated in [209] and modeled in [190], [210]–[212] to explore its effectiveness. The AC dependence of the AIFS means the MAC slot transitions are no longer completely synchronized, so modeling EDCA is more complex than modeling the DCF. In [210], the basic saturated Markov chain model of [180] is extended to different ACs by augmenting each backoff state with a long string of states representing the duration of the transmission time plus the AC-dependent AIFS. To progress to the next backoff state after any transmission, the extra string of states must be progressed through without there being another intervening transmission, either from another station, or from a different AC entity within the same station. [211] models a finite traffic load using a 3-dimensional Markov chain, with dimensions (backoff stage, backoff counter, number of packets buffered), and incorporates AC differentiation via the average duration of each backoff state. The model includes multiple transmissions within a TXOP, represented by extra states, but not virtual collisions. Correlations between transmission events are included in [212], for saturated load and without virtual collisions, by performing a fixed-point computation of the whole post-any-transmission backoff counter distribution for each AC, which provides a very accurate model at the cost of additional computational complexity. [190] models EDCA within the DSRC protocol 802.11p, so that messages are broadcast and only virtual collisions are identified by a node. Each AC is modeled by a Markov chain that uses a different probability for the channel being found busy, and when found busy, the backoff counter remains in its current state rather than decrementing. A separate Markov chain model is used for the queuing behavior, which then feeds back into the main Markov chain model as the probability of entering Idle. The models universally demonstrate the effectiveness of differentiating contention-based channel access through differentiated deferral periods. In particular, the number of lower priority class stations in the system has little impact on the throughput and delay of the higher priority class stations, which is the aim of EDCA.

3) LTE/Wi-Fi Coexistence Modeling: The modeling of LTE/Wi-Fi coexistence mechanisms has recently been an active area of research. For example, one can consider the coexistence of networks under a shared coverage area, such

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

that all nodes are within transmission range of each other. In this case, coexistence modeling and design focus on the MAC-layer and the timing of the protocol slots and frames, using Markov models.

Another modeling approach has been to focus on PHY aspects, such as transmission power, path loss, received power, SINR, and CCA power thresholds. This approach is useful for large-scale scenarios with many APs and many femtocells (being hard-connected to the core network), where nodes are no longer all within transmission range of each other. In this case, the interference plus noise power at each node impacts the channel access and density of spectral reuse. The base equation for throughput in the PHY-layer modeling approach is the Shannon capacity, $C = B\log_2(1 + \text{SINR})$, where B is the bandwidth. Rather than the protocol timing being the main focus, channel sensing becomes the dominant design feature, which is dependent on CCA thresholds and transmission powers. It is difficult to generalize because received power levels depend on the topology of the network, whereas ignoring PHY-layer considerations misses aspects of the access mechanisms and potential spectral reuse.

We compare recent literature on LTE/Wi-Fi coexistence modeling in Table VII. A comparison of coexistence schemes is also made in [213], although with little emphasis on the modeling methodology.

The interference in unmodified LTE/Wi-Fi coexistence scenarios is modeled in [214] based on the continuum field approximation and the distance from a representative UE to other eNBs and WLAN systems. The distances are modeled via a spiral approximation, to account for the UE not being centered within its cell. The Wi-Fi systems causing interference are located beyond the circle where the eNB power drops to -62 dBm, which corresponds to the Wi-Fi energy-detect carrier sensing threshold. Path loss is included, but not fading, and the throughput is given by Shannon capacity. They find that a smaller LTE cell radius reduces inter-system (LTE/Wi-Fi) interference and increases intra-system (LTE/LTE) interference.

Stochastic geometry is used in [215]. The locations of eNBs and Wi-Fi APs are modeled as independent homogeneous PPPs and then the distribution of noise and interference at a representative node is evaluated by integrating the contributions from all the other nodes in the system. The model provides the probability of supporting a given data rate, based on the Shannon capacity, and the probability of the SINR exceeding a threshold. They model LTE/Wi-Fi coexistence under unmodified LTE, discontinuous LTE transmission (LTE-U) and LBT with random backoff (LB-LBT), and show that LTE can be a good neighbor to Wi-Fi by manipulating the LTE transmission duty cycle, sensing threshold, and/or channel access priority.

A power control strategy for CSAT is proposed in [216] that controls the transmission power of each node in the network to optimize the throughput capacity, based on the SINR, using knowledge of the channel gains between all links and geometric programming.

The hyper access point (HAP) is introduced in [217] as a means of implementing stand-alone LTE-U, with both data and control are transmitted in the unlicensed band. A HAP acts as a Wi-Fi Point Coordinator, using the Wi-Fi PCF protocol to define contention free periods (CFP) for LTE access and CPs for Wi-Fi access. Using knowledge of which UE are hybrid users, having both LTE and Wi-Fi air interfaces, the HAP optimizes network utility, based on the SINR, via Nash bargaining, by choosing the CPs and CFP durations and allocating hybrid LTE/Wi-Fi users to either the CPs or CFP, while maintaining a Wi-Fi delay-quantile constraint, as modeled by [218] and [180].

In [219], CSAT is instead modeled from a MAC-layer perspective and compared to a LB-LBT variant in which, after a successful LBT procedure, a jamming signal is sent to reserve the channel until the start of the next subframe. The schemes are modeled as renewal-reward processes and the utility/throughput is maximized under a proportional fair allocation, with resource constraints included via Lagrangian multipliers. Both schemes were found to perform similarly for average on-off cycles of 20 ms and 100 ms. CSAT wastes resources on extra collisions, resulting from not performing LBT, and LB-LBT consumes resources waiting for the next subframe.

FB-LBT thoughput is modeled in [220] using Markov chains for the Wi-Fi probability of transmitting and, in turn, the distibution of the number of idle Wi-Fi MAC slots between Wi-Fi transmissions, which define the windows in which FB-LBT CCAs will succeed, assuming stationarity.

A FB-LBT variant is proposed and modeled in [221], in which the eNB jams, or reserves, the channel with a dummy packet when there is a gap in Wi-Fi transmissions shortly before each CCA. The frame period and transmission times are optimized to maximize LTE throughput, while being constrained to keep the Wi-Fi throughput and average delay no worse than if all devices were using Wi-Fi.

A number of LB-LBT works have been published based on Markov chain models. In [222], the LTE Markov chain model has two backoff stages, each with the same fixed CW. Transmissions from the first backoff stage use a high data rate and are susceptible to collision, whereas transmissions from the second backoff stage use a low data rate, based on channel quality information (CQI) feedback, and are assumed to be always successful. [223]–[225] model LB-LBT with a fixed CW, and propose load-based schemes to dynamically adapt the CW. In [223], a constrained optimization, via Lagrangian multipliers, is performed to select the CW size that minimizes the Wi-Fi collision probability while ensuring required UE data rates are met, given channel conditions, bandwidth and power bounds. An UE admission control algorithm is also offered that keeps the Wi-Fi collision probability below a threshold. In [224], the CW is found that maximizes the throughput under 'graceful coexistence', meaning individual nodes are no worse off than if all traffic were to use Wi-Fi. In [225], the model includes unsaturated LTE traffic and the CW is adapted based on a slot occupation metric. Models for LB-LBT with fixed CW and LB-LBT with exponential backoff, including average delay and unsaturated traffic are given in [226]. These Markov chain models all find that LB-LBT can lead to a suitable coexistence between LTE and Wi-Fi, and provide different feasible schemes for choosing CW sizes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE VII
COMPARISON OF RECENT LITERATURE ON LTE/WI-FI COEXISTENCE MODELING.

Access Scheme	Ref	Features / Variants
Unmodified LTE	[214], [215]	• Models are based on SINR, CCA threshold, Shannon capacity (PHY-layer) • Wi-Fi performance is significantly impacted by LTE
LTE-U/CSAT	[215]–[217], [219]	• Models are based on either: SINR/CCA (PHY-layer) or collisions/timing (MAC-layer). • Stand-alone variant: HAP. • Control variants: Tx power of each link (maximize throughput); duty-cycle (for fairness); allocations (for QoS). • If multiple eNBs present, Wi-Fi better off when eNBs use a synchronous muting pattern.
FB-LBT	[220], [221]	• Models are based on collisions/timing (MAC-layer). • Variant: pre-CCA reservation packet. • Control: frame period and Tx times (maximize LTE throughput, with Wi-Fi no worse off).
LB-LBT	[215], [219], [222]–[227]	• Models are based on either: SINR/CCA (PHY-layer) or collisions/timing (MAC-layer) • Most models assume the Wi-Fi MAC-slot structure for LB-LBT • CW Variants: fixed, adapted, exponential backoff • Adapted CW variants: maximize throughput under 'graceful coexistence'; maximize utility under proportional fair allocation; minimize Wi-Fi collision probability, while satisfying UE data rates. • Tx-power control, to optimize effective capacity or energy efficiency, given QoS and power constraints • LTE Admission control, to cap Wi-Fi collision probability • To protect Wi-Fi, LTE needs either lower channel access priority or more sensitive CCA • Models generally demonstrate that LB-LBT can achieve acceptable coexistence

The body of LTE/Wi-Fi coexistence modeling has focused on throughput and providing fairness to the incumbent Wi-Fi technology, under many variations. If LBT/CCA is used, LTE access will not be guaranteed at each attempt. The LTE latency, or duration of LTE lock-out periods, arising from LB-LBT schemes compared to FB-LBT schemes has not been modeled. Both will incur multiple consecutive failed CCA attempts under heavy channel demand. Duty cycling, without LBT, has more potential to provide deterministic latency, however, more collisions occur [219], which increase in significance as the duty cycle is shortened, as would be needed for URLLC; and some regions mandate the use of LBT. The possibility of coordinating both LTE and Wi-Fi traffic with a HAP [217], utilizing the current Wi-Fi PCF functionality, offers some promise of achieving URLLC in the unlicensed spectrum, although the PCF repetition period considered in [217] is 100 ms. The HCF additionally offers polled contention-free frames during the CP, so there is the possibility of delivering more frequent short contention-free LTE frames that would help achieve guaranteed bit rates. The use of FB-LBT, while reserving the channel when possible within a window prior to each LTE CCA [221] has the potential to reduce latency, although it comes at the cost of channel underutilization and it breaks the spirit of performing CCAs.

A new four-state semi-Markovian model is developed in [227] that quantifies the effective capacity of LAA under statistical QoS constraints. The effective capacity is the constant arrival rate that can be sustained at the input buffer while meeting a reliability constraint on the end-to-end delay (queueing plus transmission delay). A closed-form expression is derived to quantify the effective capacity of a LAA station against its QoS requirements, instantaneous transmit rate, and the numbers of LAA-BSs and Wi-Fi devices. The effective capacity is then maximized via a concave search.

The deployment scenario is important. For indoor settings, transmissions are likely to be in-range and be detected [231]. In outdoor settings, there is more variation in power levels over the cells, so there is more possibility of the transmission power and CCA threshold influencing the CCA outcomes and subsequent level of channel reuse.

C. LTE/Wi-Fi Coexistence Simulations

A comparison of simulations for different LTE/Wi-Fi coexistence schemes are given in Table VIII. Unmodified LTE, operating in the unlicensed spectrum in a simulated office environment with no mitigation techniques, such as LBT, was found to severely interfere with the incumbent Wi-Fi technology, reducing Wi-Fi channel access to below 5% of the time [228]. A comparison was made between cellular-only femtocells, dual-band femtocells operating an early FB-LBT, and a macrocell offloading to Wi-Fi via Wi-Fi hotspots, in terms of total throughput, finding that femtocells outperformed Wi-Fi offloading and dual-band femtocells outperformed cellular-only femtocells [220]. The duty-cycle approach, which uses periodic access and no channel sensing, was found in simulations to be marginally more detrimental than a FB-LBT scheme in which the eNB: senses the channel prior to each subframe until the channel is sensed available, transmits a DL frame, and then vacates the channel for a 'coexistence gap' [229]. Two schemes with subframe channel sensing were found to provide an appropriate trade-off between LTE and Wi-Fi [230]. In the first scheme, the eNB periodically senses the first (1 to 4) symbols of each subframe and, if found free, transmits for the remainder of the subframe. In the second scheme, the eNB persistently senses each subframe and when a subframe is found free, transmits for the next (1 to 4) subframes. LB-LBT is simulated with constant CW in [231], finding that, in their outdoor scenario, the LTE CCA threshold can shift the channel share balance from favoring LTE at -72 dBm to favoring Wi-Fi at -82 dBm, whereas in their indoor/outdoor scenario, with the eNBs deployed outdoors and most UE deployed indoors, there is little interference to indoor Wi-Fi networks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

The results of simulations submitted by multiple industry partners is also collated in [140] for both indoor and outdoor test scenarios. The majority of sources found at least one of their simulated LBT schemes provided fair coexistence with Wi-Fi, such that the Wi-Fi traffic was no worse off than if all traffic were supported by Wi-Fi. The particular LBT variant differed between sources though. The report recommends using LB-LBT with a CW of variable size for LAA.

D. Unlicensed Multi-carrier Access

We extend simulations that were initially performed in [232] to explore using multi-carrier access to the unlicensed spectrum as a form of multichannel diversity. The aim is to reduce the periods without access to the unlicensed spectrum. In 3GPP-Type-A multi-carrier access, only one access is allowed at a time, whereas in ETSI-Option-1, multiple backoff processes can be maintained independently.

A selection of eNB multichannel access options are considered. The eNB maintains backoff processes on between 1 and 10 unlicensed channels, and informs the UE, via the licensed channel, of which channels to monitor, waiting for a header with their ID. The eNB transmits on either:

1) Just one channel;
2) The first available channel;
3) The first available channel, while also implmenting a RTS/CTS;
4) Same as Option 3. with a less congested network.
5) All channels independently and as available; or
6) Same as option 5. with the less congested network.

For the first four options, when a transmission is made on one channel, the backoff processes on the other channels are frozen. For the last two options, the backoff processes are maintained independently of each other.

The simulations are performed to demonstrate what gains can be achieved. Each channel supports the same homogeneous traffic, comprising 10 saturated Wi-Fi stations transmitting 2 ms packets and maintaining a backoff process with CW sizes {15, 31, .., 511}. The packets are transmitted with RTS/CTS, so that in the case of a collision, the channel is occupied for 156 μs. A less congested network is also considered, in which each channel supports 5 saturated Wi-Fi stations transmitting 1 ms packets. This reduces the collision probability on each channel, without the eNB, from 0.425 to 0.328. The eNB uses LAA access priority class 1, obtaining 2 ms transmission opportunities and using CW sizes {3, 7}. When the eNB implements RTS/CTS, LTE collisions are communicated after 156 μs. To keep access fair, all devices use a DIFS for their defer periods and all devices are within transmission range. That is, all devices operating on a given unlicensed channel are located within an overlapping coverage area, such that all transmissions can be detected by the other devices operating in the channel via energy detection.

The resulting 95^{th}-quantiles of the LTE frame-start spacing (i.e., time between LTE frame starts) is plotted against the number of channels monitored in Fig. 13. The blue markers are for higher congestion, green markers for lower congestion,

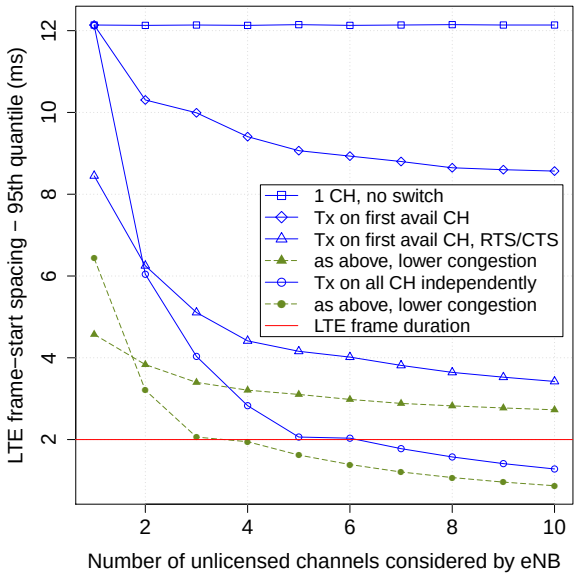


Fig. 13. Multi-carrier access to the unlicensed spectrum.

and the red line indicates the LTE frame duration. The aim is to reduce the LTE frame-start spacing to the LTE frame duration.

When the eNB monitors multiple channels (blue diamonds), the LTE frame-start spacing reduces from 12.1 ms, with one channel, down to 8.6 ms, with 10 channels. When the eNB uses a RTS/CTS mechanism (blue triangles), such that the eNB moves onto the next channel faster after a collision, all spacings are reduced, resulting in 8.5 ms spacing with one channel monitored, and 3.4 ms spacing with 10 channels monitored. When the eNB instead maintains independent backoff processes on each channel (blue circles), such that the frames on different channels can overlap, the 95^{th}-quantile of the LTE frame-start spacing reduces to 2.1 ms, with 5 channels, and to 1.3 ms with 10 channels. For the less congested network, with the eNB maintaining independent backoff processes on each channel (green circles) the 95^{th}-quantile of the spacing reduces to 2.1 ms with 3 channels and to 0.9 ms with 10 channels.

The results show that the spacing between LTE frame starts can be significantly reduced by monitoring multiple channels, achieving overlapping frames more than 95% of the time for some settings. Which channels to monitor and how many will depend on the traffic conditions. However, the results demonstrate the feasibility of utilizing the unlicensed spectrum as part of an URLLC solution.

VI. Future Research Areas

A. General Lessons and Hurdles

1) PHY-layer Lessons: To meet the critical latency requirement in URLLC, the enabling technologies are designed by reducing the TTI duration to less than 1ms. Currently, the new frame structures are discussed and standardized in 3GPP, and most research on low-latency communications focuses on how to adjust the parameters of the frame structure to support shorter TTI. The shortened waveforms provide the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

TABLE VIII
COMPARISON OF LTE/WI-FI COEXISTENCE SIMULATION LITERATURE

LTE Scheme	Ref	Finding
Unmodified LTE	[228]	LTE cripples Wi-Fi (indoor setting).
FB-LBT vs offloading	[220]	FB-LBT outperforms offloading on throughput.
FB-LBT vs duty-cycle	[229]	FB-LBT found slightly less detrimental to Wi-Fi than duty-cycle
Subframe sensing	[230]	Two channel sensing schemes offered: i) Periodically sense and Tx within each subframe; ii) Persistently sense whole subframe, then Tx for multiple subframes. Each can provide appropriate LTE/Wi-Fi trade-off (indoor setting).
LB-LBT (fixed CW)	[231]	Impact of CCA threshold on fairness: large (outdoors); little (indoors)
Unmodified/FB-LBT/LB-LBT	[140]	Recommends LB-LBT with variable-sized CW, based on multiple industry submissions for indoor and outdoor test scenarios.

potential of reducing latency substantially, without much loss in reliability, thus fulfilling the less than 1-ms TTI requirement. Furthermore, the finite blocklength information theory needs to be further developed to support high-reliability for extremely short symbols with short codes (comprising no more than 256 bits).

Since reliability is a main target in wireless communications, efforts could be put into enhancing the existing reliability-enhancing schemes, such as diversity, MCS, to support ultra-reliable. Generally, the key PHY technique to enhance reliability is diversity, which can be in the time, frequency, and/or space domains. The adaptation of MCS to channel conditions will play an important role in supporting ultra-reliable applications. Thus, a low-complexity MCS with limited feedback is desirable when there is a critical processing delay constraint.

Recent improvements in spectrum sensing and interference suppression can be utilized to reduce the sensing delay and improve the resource reuse. That means, emerging resource-reuse-based techniques, such as accurate spectrum sensing, in-band full-duplex, and grant-free NOMA, can compress the grant-request and scheduling processes, thereby reducing latency without remarkably decreasing reliability.

Accurate CSI estimation is critically challenging in URLLC because of the strictly limited delay budget. Nevertheless, many of the aforementioned PHY-layer techniques rely on accurate CSI estimation and feedback, so this fundamental technology should be thoroughly studied in URLLC.

Multiple emerging technologies can support several users in the same resource block and could enhance the capacity. Generally, grant-free NOMA can separate users multiplexed in the power domain with specific power allocation factors. Accurate CSI is not always needed in grant-free NOMA, in contrast to space division multiplexing and multiple user beamforming. Therefore, grant-free NOMA could be adopted in UL to reduce the scheduling latency, especially in massive short-packet transmissions. In addition, when space-division multiplexing and multiple-user beamforming schemes are utilized, it is common to have a large number of transmitting antennas. Therefore, the CSI estimation and feedback can have high complexity, causing an obviously greater latency in estimation and feedback processes, which is not suitable in URLLC. However, grant-free NOMA, which adopts no more than two antennas in the BS and UE, is more appropriate in the latency-limited scenario.

2) Cross-layer Lessons: Multi-connectivity creates both macro-space and network-domain diversity, leading to increased reliability without changing the latency. To be effective multi-connectivity requires coordination and synchronization between the network components, such as two APs. The multiple received signals need to be processed in combination to achieve the decoding gains.

The combined use of unlicensed and licensed bands could be considered another type of diversity. Harmonization combines protocol stacks above a certain layer, while allowing different air-interfaces to operate at the base of the stack. With harmonization at the MAC layer, scheduling for air interfaces on both the licensed and unlicensed bands could be combined in the one harmonized protocol stack, at an AP. The scheduling task is complicated due to the uncertainty of unlicensed spectrum access, but there is the potential to create diversity, thereby increasing reliability, by transmitting on both the unlicensed and licensed bands simultaneously.

In unlicensed bands, the transmit power can be dynamically adjusted to reduce the interference between adjacent subchannels and to increase the channel reuse over the network. Within licensed bands, the same applies for D2D communications, where low transmission power can be used for local communications without particularly interfering with other D2D or UE-access-point transmissions. In both cases, the use of resource management results in increased reliability of the network as a whole.

The ARQ/HARQ process has a relatively slow turn-around time, e.g., four 1-ms subframes, that exceeds the 1 ms URLLC latency target. Reducing the TTI, so that there are more than four subframes per millisecond, allows for retransmissions within the 1 ms URLLC latency budget, thereby increasing the reliability.

3) MAC-layer Lessons: There are a number of options for providing prioritized access in the licensed spectrum. For example, new URLLC bearers can be prioritized with EAB by altering the EAB categories. Then, once an URLLC bearer is established, if the bearer's activity is expected to be infrequent, such that its connection would normally be lost, there are options for providing prioritized reconnection, such as dedicated preambles. Other options include dedicating surplus resources. SPS can be used for this purpose by scheduling periodic resources. The shorter the period, the lower the latency achieved when the resources are required, but also, the more resources wasted when the resources are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

only required intermittently. Developing a transmission option that avoids the need for control signaling, while not wasting vast resources would be beneficial to URLLC, although it would likely require a probabilistic approach, such as resource pooling.

Optimizing DL allocations over a longer horizon, and then updating the optimization as newer information becomes available (e.g., on the channel state/required retransmissions/queued packets) has the potential for more efficient resource use. This could be extended to UL resource pooling, or to dynamically deciding the resources dedicated to the PRACH.

Since URLLC will likely require substantially more resources than other services, delivering URLLC may only be possible for a limited number of bearers. When a complete 5G system is settled upon, including both eMBB and URLLC bearers, the number of URLLC bearers that can be accepted without compromising the performance of all current bearers (either eMBB or URLLC) will need to be characterized, including the effects of channel conditions.

The use of D2D communications, when the end users are in close proximity, will directly reduce end-to-end latency, by reducing the number of links. The lower D2D transmission power will allow for greater spectral reuse, which may allow more URLLC bearers to be supported. It is more likely that the overlay mode will be applicable for URLLC, since the eNB can still control the links, schedule dedicated resources, and dictate the power level and MCS needed to achieve a target reliability.

In the unlicensed spectrum, Wi-Fi has the advantage of potentially fast transmissions, when the payload is small. This is considerably offset, or destroyed, by contention-based access. In contrast, the LTE frame structure and numerology set a rigid latency structure, which can be designed for high reliability, delivering upper bounds on latency with high probability.

Wi-Fi is not without QoS mechanisms. Wi-Fi QoS stations can specify their data-rate requirements. A hybrid controller (HC) can be implemented at an AP, and the HC aims to satisfy the Wi-Fi QoS stations' requirements. The HC announces periodic CFPs, during which it controls channel access by inviting stations to transmit. While not the same a scheduling, it theoretically promises the potential of much lower latency and higher reliability for Wi-Fi QoS stations.

Decentralized organization/scheduling can be achieved in the V2V broadcast scenario, which utilizes dedicated narrow channels. Proposed V2V protocols include having a locally-selected cluster head, or having each vehicle transmit its understanding of a scheduling map. The distributed protocols can provide almost contention-free access but the target transmission periodicity is 100 ms. Exploring how to reduce the periodicity, for example by utilizing multiple narrow channels with one channel for control, might make the protocols useful to unicast URLLC services more generally.

Wi-Fi also provides differentiated QoS via EDCA. EDCA defines access priority classes and provides a significant advantage, in the form of reduced congestion and shorter waiting times, to classes with higher priority. While the high-priority-user load is not too high, EDCA improves access for the high-priority users. If there are too many high-priority users though, the advantage is lost to congestion amongst the high-priority users.

Various unlicensed-spectrum access mechanisms have been explored for LTE, while considering fair LTE/Wi-Fi coexistence. Options have included duty-cycling, CSAT, FB-LBT, and LB-LBT. 3GPP has settled upon a form of LB-LBT for LAA that is very similar to EDCA. As such, there is scope for high-priority users to achieve low latency, however, this depends on the number of high-priority users, so there are no guarantees. The latency and reliability can be improved by combining accesses from multiple unlicensed bands.

B. Specific Research Areas

1) Resource Block Slicing: Diverse services should be supported in 5G within the same frequency and time radio resource blocks. Resource block slicing reserves dedicated radio resources for different services to help guarantee the interdependent latencies of different services [40]. In [233], multiplexing network slices in RANs with flexible numerology is studied and analyzed. They adopt a novel tiling concept proposed in [69], in which resource blocks are tiled into groups to allow different numerologies that in turn can be used to achieve different requirements (e.g., fast transmissions, narrow-band transmissions). A 5G virtualized RAN approach, based on not only stack (NO Stack), is proposed and evaluated in [234]–[236]. This approach supports multi-service convergence, via flexible resource block slicing, and also reduces the number of signalings and decreases latency. However, there is a trade-off between the effectiveness and efficiency of the resource block slicing, especially in URLLC services. How to mitigate the potential interruption of other types of services (such as mMTC and eMBB) by optimizing resource allocation is a topic that needs further research.

2) Advanced Signal Processing: Several techniques are proposed to achieve ultra-reliable and low latency from fundamental and theoretical perspectives, without providing directions for how to overcome realization and implementation bottlenecks. Signal processing is considered to be a vital such bottleneck. Low-complexity algorithms are studied and adopted due to the limitation of present signal processing capabilities. If parallel processing and denser processors could be simultaneously supported in a small-size electronic device, URLLC enabling techniques requiring advanced signal processing would be more likely to be realized as expected. Examples include extremely short-time channel estimation and low-delay MUD in grant-free NOMA. How to support super-fast signal processing with limited computational complexity is still an open issue. The importance of the structure of the efficient parallel/iterative process at the receiver is increasing, exceeding that of the computing power of the processors. Therefore, advanced signal detection, as well as the design of powerful electronic devices, are anticipated future research areas.

3) Location-Aware Communications: Location-aware communications are noted in [57] as one aspect of facilitating low-latency transmissions in the 5G era. In [237], authors

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials

present several challenges and solutions in designing protocol stacks that utilize location information for location-aware devices. Further, location-based physical-layer parameters, such as frame duration, physical channel, channel quality and traffic statistics, are analyzed and designed to assist low-latency, local-area, flexible, TDD systems [238], [239]. In URLLC, the accuracy and timeliness of the location information are critical, and act on each other. Therefore, further studies are needed to optimize the mechanisms to satisfy these two interrelated factors.

4) Energy Efficiency: Energy efficiency is extremely important in machine type communications, especially for sensors equipped with limited battery power and needing to operate for a couple of years. However, current research mainly focuses on the trade-off among throughput, reliability, and latency, from the perspective of spectrum efficiency, without a strong focus on energy consumption.

Thus, energy efficient URLLC enablers should also be considered, such as low power-consumption long-time spectrum sensing and accurate beamforming with high-order space diversity and huge power gains. Thus, the design of energy efficient URLLC enablers, such as low power-consumption long-time spectrum sensing and accurate beamforming with high-order space diversity and huge power gains, is an emerging research area.

5) Multichannel Diversity: As demonstrated in Section V-D, there is potential to utilize the Type-A multi-carrier access options in the unlicensed spectrum, maintaining separate channel-access backoff processes for each carrier, so as to achieve multichannel diversity. The idea is to transmit with the earliest access opportunity, thus reducing the latency. Fast frequency hopping techniques could then be used to jump between available carriers.

The reliability for a given latency is affected by the number of unlicensed channels being monitored and their traffic profiles. By maintaining a stochastic model for the traffic on each channel, the theory of optimal stopping point (OSP) [240] could be applied to decide when to switch channels. OSP selects a channel switching time so as to maximize a reward function, based on the stochastic model and channel observations, which could be specified to either minimize latency or maximize reliability, Developing a model that characterizes the traffic of a given channel with implementable computation complexity is a challenge. Converting a traffic model into the access-delay distribution, again with implementable computation complexity, is another challenge.

6) Carrier Aggregation Scheduling: To implement multi-channel diversity, carrier aggregation resource allocation and scheduling is required, which is surveyed in [241]. They note that UE need to estimate and report the channel quality of each component carrier (CC). To then take full advantage of CQI and achieve spectral efficiency, joint CC and RB scheduling is needed. Efficient schedulers are typically characterized by higher delays, whereas simpler schedulers waste resources. For URLLC, low latency is needed as well as the controlled reliability offered by the CQI, so research into more efficient, yet fast, cross-channel schedulers are needed. An autonomous channel bonding method is proposed in [242] that allows the quick and efficient selection of unlicensed channels for aggregation, while relying on limited feedback.

7) Unlicensed Channel Profiling: Q-learning is applied in [243] to predict the expected throughput from an access attempt on each of multiple unlicensed channels, under LTE-U (ETSI FB-LBT), and to then choose the best channel. Q-learning randomly explores the available channels, favoring those with higher expected throughput, and applying a 'cooling' factor so that channel exploration decreases over time. With multiple small cells applying Q-learning, their channel selections self-organize over time and adapt to changes in channel congestion. Distributed heuristically accelerated Q-learning is applied in [244] to achieve flexible dynamic spectrum access (DSA), while avoiding inter-cell interference between adjacent cells; in this case, adjacent eNB's exchange bit maps of the resources they are using. Channel profiling and the development of algorithms to select the best unlicensed channel for URLLC, are potential research topics.

VII. Conclusion

We have surveyed the current PHY layer and MAC layer mechanisms that reduce latency or improve reliability in communications systems. We have then identified ones that are most relevant to help enable URLLC in both the licensed spectrum and the unlicensed spectrum.

Some PHY mechanisms are applicable to both licensed spectrum and unlicensed spectrum, such as shortening the TTI and allowing a more flexible frame structure, whereas other mechanisms are more suitable to either the licensed spectrum or unlicensed spectrum, such as frequency hopping.

MAC layer mechanisms in the licensed spectrum include streamlining high priority bearers, reducing control signaling delays, and reducing the number of links. In the unlicensed spectrum, gains in latency and reliability are made by reducing collisions; providing prioritized, yet still contention-based, access processes; and by attempting to provide centrally controlled transmission opportunities. Mechanisms exist within the Wi-Fi protocols to allow some level of central coordination in the unlicensed spectrum. However, these mechanisms rely on the Wi-Fi coordination control mechanisms, so are not presently available to LTE operating in the unlicensed spectrum. A critical challenge is to utilize the unlicensed spectrum so as to provide low-latency, high-reliability communications, possibly providing guaranteed bit rates, while not significantly impacting on a particular technology's right to also use the channel.

A promising MAC-layer direction for utilizing the unlicensed spectrum as part of an URLLC solution, is being able to predict when the unlicensed spectrum will be available, so that schedulers can be informed.

Acknowledgment

This work was supported by the Huawei Innovation Research Program (No. HIRPO20160105).

REFERENCES

[1] 3GPP TR 38.913 V14.2.0 Study on scenarios and requirements for next generation access technologies (Release 14).

[2] H. V. K. Mendis and F. Y. Li, "Achieving ultra reliable communication in 5G networks: A dependability perspective availability analysis in the space domain," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2057-2060, Sep. 2017.

[3] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, and A. Puschmann, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70-78, 2017.

[4] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "The 5G-enabled tactile internet: Applications, requirements, and architecture," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2016, pp. 61-66.

[5] D. Bruemmer. (2016, Sep.) Self-driving cars need to be more reliable. [Online]. Available: http://www.roboticstrends.com/article/self_driving_cars_need_to_be_more_reliable

[6] L. Eldada. (2014, Mar.) Today's LiDARs and GPUs enable ultra-accurate GPS-free navigation with affordable SLAM. [Online]. Available: http://on-demand.gputechconf.com/gtc/2014/presentations/S4761-lidar-ultra-accurate-affordable-localization-mapping.pdf

[7] O. N. C. Yilmaz, "Ultra-reliable and low-latency 5G communication," in *Proc. European Conference on Networks and Communications (EuCNC)*, 2016.

[8] A. Rahmani, N. Thanigaivelan, T. Gia, J. Granados, B. Negash, P. Liljeberg, and H. Tenhunen, "Smart e-Health gateway: Bringing intelligence to internet-of-things based ubiquitous healthcare systems," in *Proc. IEEE Consumer Communications and Networking Conference (CCNC)*, Jan. 2015, pp. 9-12.

[9] O. Kocabaş, and T. Soyata, "E-Health and telemedicine: Concepts, methodologies, tools, and applications," Information Resources Management Association (USA), Sep. 2015.

[10] Y. M. Huang, M. Y. Hsieh, H. C. Chao, S. H. Hung, and J. H. Park, "Pervasive, secure access to a hierarchical sensor-based healthcare monitoring architecture in wireless heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 4, pp. 400-411, May 2009.

[11] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators and virtual environments*, vol. 6, no. 4, pp. 355-385, 1997.

[12] F. Zheng, T. Whitted, A. Lastra, P. Lincoln, A. State, A. Maimone, and H. Fuchs, "Minimizing latency for augmented reality displays: Frames considered harmful," in *Proc. IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*, Munich, Germany, Sep. 2014, pp. 195-200.

[13] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Netw.*, vol. 32, no. 2, pp. 78-84, Mar. 2018.

[14] H. Flatt, N. Koch, C. Röcker, A. Günter and J. Jasperneite, "A context-aware assistance system for maintenance applications in smart factories based on augmented reality and indoor localization," in *Proc. IEEE Conference on Emerging Technologies & Factory Automation (ETFA)*, Luxembourg, Sep. 2015, pp. 1-4.

[15] S.-F. Persa, "Sensor fusion in head pose tracking for augmented reality," Institutional Repository, Jun. 2016.

[16] K. Trichias, "Modeling and evaluation of LTE in intelligent transportation systems," University of Twente, Oct. 2011.

[17] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen, and P. Mogensen, "Automation for on-road vehicles: Use cases and requirements for radio design," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2015, pp. 1-5.

[18] ICT-317669 METIS project, "Scenarios, requirements and KPIs for 5G mobile and wireless system," Deliverable D1.1, Apr. 2013.

[19] X. Yang, L. Liu, N. H. Vaidya, and F. Zhao, "A vehicle-to-vehicle communication protocol for cooperative collision warning," *Mobile and Ubiquitous Systems: Networking and Services*, Aug. 2004, pp. 114-123.

[20] F. Fitzek, and G. Fettweis. (2014 Sep.) Agile cloud supporting the tactile internet. [Online]. Available: http://www.ict-ijoin.eu/wp-content/uploads/2015/03/7a_Fitzek_Agile-Cloud.pdf

[21] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460-473, 2016.

[22] Z. Shi, H. Zou, M. Rank, L. Chen, S. Hirche, and H. J. Mueller, "Effects of packet loss and latency on temporal discrimination of visual-haptic events," *IEEE Trans. Haptics*, vol. 3, no. 1, pp. 28-36, 2009.

[23] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Netw.*, vol. 32, no. 2, pp. 24-31, Mar. 2018.

[24] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1391-1396.

[25] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st Int. Conf. on 5G for Ubiquitous Connectivity*, Dec. 2014, pp. 146-151.

[26] E. G. Ström, P. Popovski, and J. Sachs, "5G ultra-reliable vehicular communication," *Mathematics*, vol. 14, pp. 204-217, 2015.

[27] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1184-1189.

[28] 3GPP TR 22.862 V14.1.0 Feasibility study on new services and markets technology enablers for critical communications (Release 14), Oct. 2016.

[29] 3GPP TR 38.801 V2.0.0 Study on new radio access technology; radio access architecture and interfaces (Release 14).

[30] 3GPP TR 38.804 V14.0.0 Study on new radio access technology; radio interface protocol aspects (Release 14).

[31] 3GPP TR 38.912 V14.0.0 Study on new radio (NR) access technology (Release 14).

[32] 3GPP TR 38.802 V14.2.0 Study on new radio access technology physical layer aspects (Release 14), Sep. 2017.

[33] A. Goldsmith. *Wireless communications*. Cambridge university press, 2005.

[34] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232-4265, Jul. 2014.

[35] Y. Hu, A. Schmeink, and J. Gross, "Blocklength-limited performance of relaying under quasi-static rayleigh channels," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4548-4558, Jul. 2016.

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

[37] A. Gohil, H. Modi, and S. K. Patel, "5G technology of mobile communication: A survey," in *Proc. Int. Conf. Intelligent Systems and Signal Processing (ISSP)*, Mar. 2013, pp. 288-292.

[38] A. Gupta, and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206-1232, 2015.

[39] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617-1655, Third Quart. 2016.

[40] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, pp. 1-7, 2018.

[41] Z. Ma, Z. Zhang, Z. Ding, P. Fan, and H. Li, "Key techniques for 5G wireless communications: network architecture, physical layer, and MAC layer perspectives," *Science China information sciences*, vol. 58, no. 4, pp. 1-20, 2015.

[42] A. Nasrallah, A. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ULL) networks: A comprehensive survey covering the IEEE TSN standard and related ULL research," *arXiv preprint*, arXiv:1803.07673, 2018.

[43] P. Heise, N. Tobeck, O. Hanka, and S. Schneele, "SAFDX: deterministic high-availability ring for industrial low-cost networks," in *Proc. 7th International Workshop on Communication Technologies for Vehicles (Nets4Cars-Fall)*, Oct. 2014, pp. 40-44.

[44] A. F. Cattoni, D. Chandramouli, C. S. Rainer, and S. P. Zanier, "Mobile low latency services in 5G," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, 2015.

[45] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges," *Wireless Networks*, vol. 21, no. 8, pp. 2657-2676, 2015.

[46] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347-2376, Fourth Quart. 2015.

[47] X. Zhang, T. Lv, W. Ni, J. M. Cioffi, N. C. Beaulieu, and Y. J. Guo, "Energy-efficient caching for scalable videos in heterogeneous networks," *IEEE J. Select. Areas Comm.*, accepted on 16/04/2018.

[48] X. Lyu, Wei Ni, H. Tian, R. P. Liu, X. Wang, G. B. Giannakis, and A. Paulraj, "Optimal schedule of mobile edge computing for internet of things using partial information," *IEEE J. Sel. Areas Comm.*, vol. 35, no. 11, pp. 2606 - 2615, Nov. 2017.

[49] X. Lyu, C. Ren, W. Ni, H. Tian, and R. P. Liu, "Distributed optimization of collaborative regions in large-scale inhomogeneous fog computing," *IEEE J. Sel. Areas Comm.*, accepted on 27/02/2018.

[50] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang and R. P. Liu, "Energy-efficient admission of delay-sensitive tasks for mobile edge computing," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2603-2616, Jun. 2018.

[51] S. Lin, J. Yu, W. Ni, and R. P. Liu, "Decoupling 5G network control: Centralized coordination and distributed adaptation," *Int'l J. Comput. Comm. Control (IJCCC)*, accepted on 21/02/2018.

[52] S. Zou, J. Tang, W. Ni, R. P. Liu, and L. Wang, "Resource multi-objective mapping algorithm based on virtualized network functions: RMMA," *Applied Soft Computing*, vol. 66, pp. 220 - 231, May 2018.

[53] X. Lyu, H. Tian, W. Ni, R. P. Liu, and P. Zhang, "Adaptive centralized clustering framework for software-defined wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8553 - 8557, Sep. 2017.

[54] X. Chen, W. Ni, I. B. Collings, X. Wang, and S. Xu, "Distributed placement and online optimization of virtual machines for network service chains," accepted to *ICC 2018*.

[55] A. A. Khan, M. Abolhasan, and W. Ni, "5G next generation VANETs using SDN and fog computing framework," in *Proc. IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, 2018, pp. 1-6.

[56] X. Chen, W. Ni, T. Chen, I. B. Collings, X. Wang, R. P. Liu, and G. B. Giannakis, "Distributed stochastic optimization of network function virtualization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1-6.

[57] 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.*, pp. 1-34, 2018.

[58] Motorola, "Motorola solutions white paper: Public safety LTE, prioritization and preemption, quality of service," Motorola Solutions, 2016.

[59] ETSI EN 300 392-2 v3.4.1, Terrestrial trunked radio (TETRA); voice plus data (V+D); Part 2: air interface (AI), Aug. 2010.

[60] E. Guttman and A. Scrase, "Status and progress on mobile critical communications standards," presented at *CCE 2017*, Copenhagen, DK, Feb. 8-9, 2017

[61] 3GPP TS 22.179 v14.3.0, Mission critical push to talk (MCPTT) over LTE; Stage 1 (Release 14), Dec. 2016.

[62] 3GPP TS 23.281 v14.0.0, Functional architecture and information flows to support mission critical video (MCVideo); Stage 2 (Release 14), Dec. 2016.

[63] 3GPP TS 22.468 v14.0.0 Group communication system enablers for LTE (GCSE_LTE) (Release 14), Mar. 2017.

[64] 3GPP TR 36.881 V14.0.0 Study on latency reduction techniques for LTE (Release 14), Jun. 2016.

[65] S. A. Ashraf, F. Lindqvist, R. Baldemair, and B. Lindoff, "Control channel design trade-offs for ultra-reliable and low-latency communication system," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1-6.

[66] M. Ibrahim and W. Xu, "A comparison of symbol-wise and self-contained frame structure for 5G services," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1-6.

[67] T. Levanen, J. Pirskanen, and M. Valkama, "Radio interface design for ultra-low latency millimeter-wave communications in 5G era," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1420-1426.

[68] A. Damnjanovic, W. Chen, S. Patel, Y. Xue, and K. Hosseini, "Techniques for enabling low latency operation in LTE networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1-7.

[69] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53-59, 2016.

[70] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1-6.

[71] Y. Liu et al., "Waveform design for 5G networks: Analysis and comparison," *IEEE Access*, pp. 1-9, 2017.

[72] X. Huang, J. A. Zhang, and Y. J. Guo, "Out-of-band emission reduction and a unified framework for precoded OFDM," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 151-159, 2015.

[73] B. Farhang-Boroujeny and H. Moradi, "OFDM inspired waveforms for 5G," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2474-2492, Fourth Quart. 2016.

[74] V. Vakilian, T. Wild, F. Schaich, S. Brink, and J. Frigon, "Universal-filtered multi-carrier technique for wireless systems beyond LTE," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 1-6.

[75] X. Zhang, L. Chen, J. Qiu, and J. Abdoli, "On the waveform for 5G," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 74-80, 2016.

[76] G. Berardinelli, K. I. Pedersen, T. B. Sorensen, and P. Mogensen, "Generalized DFT-spread-OFDM as 5G waveform," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 99-105, 2016.

[77] A. Sahin, R. Yang, E. Bala, M. C. Beluri, and R. L. Olesen, "Flexible DFT-S-OFDM: Solutions and challenges," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 106-112, 2016.

[78] I. Chih-Lin, S. Han, Z. Xu, S. Wang, Q. Sun, and Y. Chen, "New paradigm of 5G wireless internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 474-482, 2016.

[79] Z. E. Ankarali, B. Pekoz, and H. Arslan, "Flexible radio access beyond 5G: A future projection on waveform, numerology & frame design principles," *IEEE Access*, pp. 1-16, 2017.

[80] J. Yli-Kaakinen et al., "Efficient fast-convolution based waveform processing for 5G physical layer," *IEEE J. Sel. Areas Commun.*, pp. 1-19, 2017.

[81] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. of the 18th ACM Int. Conf. on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2015, pp. 1-10.

[82] J. Östman, G. Durisi, E. G. Ström, J. Li, H. Sahlin, and G. Liva, "Low-latency ultra reliable 5G communications: Finite-blocklength bounds and coding schemes," in *Proc. IEEE Conf. on Systems, Communications and Coding*, Feb. 2017, pp. 1-7.

[83] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618-629, 2016.

[84] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711-1726, 2016.

[85] D. Ohmann, M. Simsek, and G. P. Fettweis, "Achieving high availability in wireless networks by an optimal number of Rayleigh-fading links," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1402-1407.

[86] N. Brahmi, O. N. C. Yilmaz, K. W. Helmersson, S. A. Ashraf, and J. Torsner, "Deployment strategies for ultra-reliable and low-latency communication in factory automation," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1-6.

[87] Y. Gao, T. Yang, and B. Hu, "Improving the transmission reliability in smart factory through spatial diversity with ARQ," in *Proc. IEEE/CIC Int. Conf. on Commun. in China (ICCC)*, Jul. 2016, pp. 1-5.

[88] C. She, C. Yang, and T. Q. S. Quek, "Uplink transmission design with massive machine type devices in tactile internet," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1-6.

[89] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, M. Latva-aho, and C. S. Hong, "Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2041-2044, 2017.

[90] S. R. Panigrahi, N. Bjorsell, and M. Bengtsson, "Feasibility of large antenna arrays towards low latency ultra reliable communication," in *Proc. IEEE Int. Conf. on Industrial Technology (ICIT)*, Mar. 2017, pp. 1289-1294.

[91] A. Tassi, I. Chatzigeorgiou, and D. E. Lucani, "Analysis and optimization of sparse random linear network coding for reliable multicast services," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 285-299, 2016.

[92] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jantti, "Link adaptation design for ultra-reliable communications," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, May 2016, pp. 1-5.

[93] B. Farayev, Y. Sadi, and S. C. Ergen, "Optimal power control and rate adaptation for ultra-reliable M2M control applications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1-6.

[94] F. Schuh and J. B. Huber, "Punctured vs. multidimensional TCM - A comparison w.r.t. complexity," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1408-1413.

[95] U. Oruthota, F. Ahmed, and O. Tirkkonen, "Ultra-reliable link adaptation for downlink MISO transmission in 5G cellular networks," *Information*, vol. 7, no. 1, pp. 1-18, 2016.

[96] G. Liva, L. Gaudio, and T. Ninacs, "Code design for short blocks?: A survey," in *Proc. European Conf. on Networks and Communications Workshops (EuCNC Wkshps)*, 2016, pp. 1-5.

[97] Shirvanimoghaddam, Mahyar, Yonghui Li, and Branka Vucetic. "Near-capacity adaptive analog fountain codes for wireless channels," *IEEE Commun. Lett.* pp. 2241-2244, 2013.

[98] Y. Hori and H. Ochiai, "A low PAPR subcarrier hopping multiple access with coded OFDM for low latency wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1-6.

[99] Y. Hori and H. Ochiai, "A design of multiuser detection and decoding for subcarrier hopping multiple access based on coded OFDM," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1-6.

[100] S. Xia, X. Han, X. Yan, Z. Zuo, and F. Bi, "Uplink control channel design for 5G ultra-low latency communication," in *Proc. IEEE 27th Annu. Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1-6.

[101] S. Saponara, F. Giannetti, B. Neri, and G. Anastasi, "Exploiting mm-wave communications to boost the performance of industrial wireless networks," *IEEE Trans. Ind. Informatics*, vol. 13, no. 2, pp. 1-10, 2017.

[102] P. K. Sahoo and D. Sahoo, "Sequence-based channel hopping algorithms for dynamic spectrum sharing in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2814-2828, 2016.

[103] A. Ali and W. Hamouda, "Advances on Spectrum Sensing for Cognitive Radio Networks: Theory and Applications," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1277-1304, Second Quart. 2017.

[104] K. Cichoń, A. Kliks and H. Bogucka, "Energy-Efficient Cooperative Spectrum Sensing: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1861-1886, Third Quart. 2016.

[105] Z. Zeinalkhani and A. H. Banihashemi, "Ultra low-complexity detection of spectrum holes in compressed wideband spectrum sensing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1-7.

[106] B. Shahrasbi, N. Rahnavard, and A. Vosoughi, "Cluster-CMSS: A cluster-based coordinated spectrum sensing in geographically dispersed mobile cognitive radio networks," *IEEE Trans. Veh. Technol.*, pp. 1-9, 2016.

[107] Y. Ma, Y. Gao, Y.-C. Liang, and S. Cui, "Reliable and efficient sub-nyquist wideband spectrum sensing in cooperative cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2750-2762, 2016.

[108] X. Huang, S. Member, Y. Xu, J. Wu, S. Member, and W. Zhang, "Non-cooperative spectrum sensing with historical sensing data mining in cognitive radio," *IEEE Trans. Veh. Technol.*, pp. 1-9, 2017.

[109] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637-1652, 2014.

[110] D. Kim, H. Lee and D. Hong, "A survey of in-band full-duplex transmission: From the perspective of PHY and MAC layers," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2017-2046, Fourth Quart. 2015.

[111] W. Afifi and M. Krunz, "Incorporating self-interference suppression for full-duplex operation in opportunistic spectrum access systems," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 4, pp. 2180-2191, 2015.

[112] W. Afifi and M. Krunz, "TSRA: An adaptive mechanism for switching between communication modes in full-duplex opportunistic spectrum access systems," *IEEE Trans. Mob. Comput.*, vol. 15, no. 8, pp. 1536-1233, 2016.

[113] A.-A. Boulogeorgos, H. Bany Salameh, and G. Karagiannidis, "Spectrum sensing in full-duplex cognitive radio networks under hardware imperfections," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2072-2084, 2017.

[114] A. Yadav, O. A. Dobre, and N. Ansari, "Energy and traffic aware full-duplex communications for 5G systems," *IEEE Access*, vol. 5, pp. 11278-11290, 2017.

[115] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tutorials*, pp. 1-42, 2016.

[116] J. Zeng et al., "Investigation on Evolving Single-Carrier NOMA Into Multi-Carrier NOMA in 5G." *IEEE Access*, vol. 6, pp. 48268-48288, 2018.

[117] Z. Ding et al., "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185-191, Feb. 2017.

[118] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[119] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294-2323, Third Quart. 2018.

[120] M. Mohammadkarimi, M. A. Raza and O. A. Dobre, "Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 40-50, Dec. 2018.

[121] A. Yadav and O. A. Dobre, "All technologies work together for good: A glance at future mobile networks," *IEEE Wireless Commun.*, vol. 25, no. 4, pp. 10-16, Aug. 2018.

[122] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the internet of things," *IEEE Access*, vol. 4, pp. 1393-1405, 2016.

[123] L. Yang, Y. Liu, and Y. Siu, "Low complexity message passing algorithm for SCMA system," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2466-2469, 2016.

[124] Z. Zhang, X. Wang, Y. Zhang, and Y. Chen, "Grant-free rateless multiple access: A novel massive access scheme for internet of things," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2019-2022, 2016.

[125] J. Zhang et al., "PoC of SCMA-based uplink grant-free transmission in UCNC for 5G," *IEEE J. Sel. Areas Commun.*, pp. 1-10, 2017.

[126] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320-2323, 2016.

[127] Y. Du et al., "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812-2828, 2017.

[128] K. Wu, W. Ni, T. Su, and R. P. Liu, and Y. J. Guo, "Fast and accurate estimation of angle-of-arrival for satellite-borne wideband communication system," *IEEE J. Select. Areas Comm.*, vol. 36, no. 2, pp. 314 - 326, Feb. 2018.

[129] K. Wu, W. Ni, T. Su, R. P. Liu, and J. Y. Guo, "Robust unambiguous estimation of angle-of-arrival in hybrid array with localized analog subarrays," *IEEE Trans. Wireless Comm.*, accepted on 07/02/2018.

[130] M. Wang, J. Cai, F. Tseng, and C. Hsu, "A low-complexity 2-D angle of arrival estimation in massive MIMO systems," in *Proc. Int. Computer Symp. (ICS)*, Dec. 2016, pp. 710-713.

[131] J. A. Zhang, W. Ni, P. Cheng, and Y. Lu, "Angle-of-arrival estimation using different phase shifts across subarrays in localized hybrid arrays," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2205-2208, 2016.

[132] M. Serror, C. Dombrowski, K. Wehrle, and J. Gross, "Channel coding versus cooperative ARQ: Reducing outage probability in ultra-low latency wireless communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1-6.

[133] H. Shariatmadari, S. Iraji, and R. Jantti, "Analysis of transmission methods for ultra-reliable communications," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2015, pp. 2303-2308.

[134] B. S. Sheshachalam, S. Kalyanasundaram, and S. K. V, "A novel HARQ Pooling scheme for improved multi-connectivity in 5G cloud RAN," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1-7.

[135] H. Shariatmadari, S. Iraji, Z. Li, M. A. Uusitalo, and R. Jäntti, "Optimized transmission and resource allocation strategies for ultra-reliable communications," in *Proc. IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Dec. 2016, pp. 1-6.

[136] P. Marsch, I. Da Silva, Ö. Bulakci, M. Tesanovic, S. E. El Ayoubi, and M. Säily, "Emerging network architecture and functional design considerations for 5G radio access," *Trans. Emerging Tel. Technol.*, vol. 27, no. 9, pp. 1168-1177, 2016.

[137] IEEE std 802.11-2007, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specification," IEEE Std 802.11-2007, IEEE, Jun. 2007.

[138] W. Sun, E. G. Strom, F. Brannstrom, Y. Sui, and K. C. Sou, "D2D-based V2V communications with latency and reliability constraints," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1414-1419.

[139] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer transmission design for tactile internet," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1-6.

[140] 3GPP TR 36.889 v13.0.0. Study on licensed-assisted access to unlicensed spectrum (Release 13), Jun. 2015.

[141] Z. Nan, T. Chen, X. Wang, and W. Ni, "Energy-efficient transmission schedule for delay-limited bursty data arrivals under nonideal circuit power consumption," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6588-6600, 2016.

[142] A. He, L. Wang, Y. Chen, K. Wong, and M. Elkashlan, "SE and EE of uplink D2D underlaid massive MIMO cellular networks with power control," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1-6.

[143] A. Awad, A. Mohamed, and C. Chiasserini, "User-centric network selection in multi-RAT systems," in *Proc. IEEE Wireless Commun. Netw. Conf. workshops (WCNC Wkshps)*, Apr. 2016, pp. 1-6.

[144] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5G intra-frequency networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1-6.

[145] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Evaluation of adaptive active set management for multi-connectivity in intra-frequency 5G networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2016, pp. 1-6.

[146] K. M. S. Huq et al., "Enhanced C-RAN using D2D network," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 100-107, 2017.

[147] A. Orsino et al., "Effects of heterogeneous mobility on D2D- and drone-assisted mission-critical MTC in 5G," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 79-87, 2017.

[148] C. Kilinc, J. F. Monserrat, M. C. Filippou, N. Kuruvatti, A. A. Zaidi, I. Da Silva, and M. Mezzavilla, "New radio 5G user plane design alternatives: One 5G air interface framework supporting multiple services and bands," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1-6.

[149] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86-93, 2013.

[150] Y. C. Pang, S. L. Chao, G. Y. Lin, and H. Y. Wei, "Network access for M2M/H2H hybrid systems: a game theoretic approach," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 845-848, 2014.

[151] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10-17, 2015.

[152] R. G. Cheng, J. Chen, D. W. Chen, and C. H. Wei, "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 6, pp. 2956-2968, 2015.

[153] ICT-671680 METIS-II, Deliverable 2.2, "Draft overall 5G RAN design," Jun. 2016.

[154] M. Centenaro, L. Vangelista, S. Saur, A. Weber, and V. Braun, "Comparison of collision-free and contention-based radio access protocols for the internet of things," *IEEE Trans. Commun.*, 2017.

[155] 3GPP TS 23.203 v14.3.0, Policy and charging control architecture (Release 14), Mar. 2017.

[156] S. Y. Lien, C. C. Chien, G. S. T. Liu, H. L. Tsai, R. Li, and Y. J. Wang, "Enhanced LTE device-to-device proximity services," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 174-182, 2016.

[157] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Introduction to ultra reliable and low latency communications in 5G," arXiv preprint arXiv:1704.05565, 2017.

[158] 3GPP TS 36.321 v13.2.0 (2016-06), Medium access control (MAC) protocol specification (Release 13), Jun. 2016.

[159] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678-700, Second Quart. 2013.

[160] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150-154, Feb. 2001.

[161] R. Basukala, H. Mohd Ramli, and K. Sandrasegaran, "Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system," in *Proc. AH-ICI, Kathmundu, Nepal*, Nov. 2009, pp. 1-5.

[162] G. Piro, L. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-level downlink scheduling for real-time multimedia services in LTE networks," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1052-1065, Oct. 2011.

[163] N. Abu-Ali, A. E. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in LTE and LTE-advanced: Tutorial, survey and evaluation framework," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1239-1265, Third Quart. 2014

[164] M. A. Mehaseb, Y. Gadallah, A. Elhamy, and H. Elhennawy, "Classification of LTE uplink scheduling techniques: An M2M perspective," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1310-1335, Second Quart. 2016.

[165] M. Kalil, A. Shami, and A. Al-Dweik, "QoS-aware power-efficient scheduler for LTE uplink," *IEEE Trans. on Mob. Comp.*, vol. 14, no. 8, pp. 1672-1685, 2015.

[166] J. Liu, C. Hu, Z. Ma, K. Zheng, and W. Wang, "Semi-persistent scheduling for VoIP service in the LTE-Advanced relaying networks," in *Proc. International Conf. on Commun., Circuits and Systems (ICCCAS)* , Jul. 2010, pp. 54-58.

[167] M. R. Tabany, C. G. Guy, and R. S. Sherratt, "A novel downlink semi-persistent packet scheduling scheme for VoLTE traffic over heterogeneous wireless networks," *EURASIP Journal on Wireless Communications and Networking*, 2017:62, 2017.

[168] C. Hoymann et al., "LTE release 14 outlook," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 44-49, 2016.

[169] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson, and J. Ansari, "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G," in *Proc. IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sep. 2016, pp. 1-8.

[170] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1-5.

[171] J. Brown, and J. Y. Khan, "A predictive resource allocation algorithm in the LTE uplink for event based M2M applications," *IEEE Trans. Mobile Comput.*, vol. 14, no. 12, pp. 2433-2446, 2015.

[172] M. Yang, and T. Chin, Qualcomm Inc., "Service request, scheduling request, and allocation of radio resources for service contexts," U.S. Patent No. US 2017/0094654 A1, 30 Mar., 2017.

[173] X. Lin, J. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40-48, 2014.

[174] S. Y. Lien, C. C. Chien, F. M. Tseng, and T. C. Ho, "3GPP device-to-device communications for beyond 4G cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 29-35, 2016.

[175] S. M. Lopez. (2016, Jun.) An overview of D2D in 3GPP LTE standard, OrangeTM. [Online]. Available: http://d2d-4-5g.gforge.inria.fr/Workshop-June2016/slides/Overview_LTE_D2D.pdf

[176] 3GPP TS 36.213 v14.1.0, Physical layer procedures (Release 14), Dec 2016.

[177] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1923-1940, Fourth Quart. 2015.

[178] Y. Xu, R. Yin, T. Han, and G. Yu, "Dynamic resource allocation for device-to-device communication underlaying cellular networks," *International Journal of Communication Systems*, vol. 27, no. 10, pp. 2408-2425, 2014.

[179] K. Xu, M. Gerla, and S. Bae, "Effectiveness of RTS/CTS handshake in IEEE 802.11 based ad hoc networks," *Ad hoc networks*, vol. 1, no. 1, pp. 107-123, 2003.

[180] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535-547, 2000.

[181] N. Sai Shankar, D. Dash, H. El Madi, G. Gopalakrishnan "WiGig and IEEE 802.11ad for multi-gigabyte-per-second WPAN and WLAN," submitted to Cornell University Library 2012 cited as arXiv:1211.7356.

[182] M.X. Gong, R. Stacey, D. Akhmetov, and S. Mao, "A directional CSMA/CA protocol for mmWave wireless PANs," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2010, pp. 1-6.

[183] Y. B. Ko, V. Shankarkumar, and N. H. Vaidya, "Medium access control protocols using directional antennas in ad hoc networks," *IEEE International Conference on Computer Communications (INFOCOM)*, vol. 1, pp. 13-21, 2000.

[184] G. H. Sim, T. Nitsche, and J. C. Widmer, "Addressing MAC layer inefficiency and deafness of IEEE802.11ad millimeter wave networks using a multi-band approach," in *Proc. IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2016, pp. 1-6.

[185] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2377-2396, Fourth Quart. 2015.

[186] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of DSRC and cellular network technologies for V2X communications: A survey," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9457-9470, 2016.

[187] IEEE Std 1609.4-2106, IEEE Standard for Wireless Access in Vehicular Environments (WAVE)-Multi-Channel Operation, 2016.

[188] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162-1182, 2011.

[189] F. Cunha, L. Villas, A. Boukerche, G. Maia, A. Viana, R. A. Mini, and A. A. Loureiro, "Data communication in VANETs: Protocols, applications and challenges," *Ad Hoc Networks*, vol. 44, pp. 90-103, 2016.

[190] Y. Yao, L. Rao, and X. Liu, "Performance and reliability analysis of IEEE 802.11p safety communication in a highway environment," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4198-4212, 2013.

[191] Y. P. Fallah, C. L. Huang, R. Sengupta, and H. Krishnan, "Analysis of information dissemination in vehicular ad-hoc networks with application to cooperative vehicle safety systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 1, pp. 233-247, 2011.

[192] K. A. Hafeez, L. Zhao, B. Ma, and J. W. Mark, "Performance analysis and enhancement of the DSRC for VANET's safety applications," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3069-3083, 2013.

[193] K. A. Hafeez, A. Anpalagan, and L. Zhao, "Optimizing the control channel interval of the DSRC for vehicular safety applications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3377-3388, 2016.

[194] X. Zhang, H. Su, and H. H. Chen, "Cluster-based multi-channel communications protocols in vehicle ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, 2006

[195] N. Lu, Y. Ji, F. Liu, and X. Wang, "A dedicated multi-channel MAC protocol design for VANET with adaptive broadcasting," in *Proc. Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2010, pp. 1-6.

[196] F. Borgonovo, A. Capone, M. Cesana, and L. Fratta, "RR-ALOHA, a reliable R-ALOHA broadcast channel for ad-hoc inter-vehicle communication networks," in *Proc. of Med-Hoc-Net*, 2002.

[197] C. Campolo, A. Molinaro, G. Araniti, and A. O. Berthet, "Better platooning control toward autonomous driving: An LTE device-to-device communications strategy that meets ultralow latency requirements," *IEEE Veh. Technol. Mag.*, vol. 12, no. 1, pp. 30-38, 2017.

[198] Draft ETSI EN 301 893 v2.0.7, "5 GHz RLAN; Harmonised standard covering the essential requirements of article 3.2 of Directive 2014/53/EU," Nov. 2016.

[199] J. Liu, W. Xiao, and Y. Xia, Y., "Device, network, and method for communications with opportunistic transmission and reception," U.S. Patent No. US 2017/0142751 A1, 18 May, 2017.

[200] "MulteFireTM technology progress and benefits, and how it enables a new breed of neutral hosts," slides, May 24, 2016.

[201] MulteFire Release 1.0, Technical Paper, "A New Way to Wireless," $\mathrm{Multefire}^{\mathrm{TM}}$, 2017.

[202] Y. Wang, J. Xu, and L. Jiang, "Challenges of system-Level simulations and performance evaluation for 5G wireless networks," *IEEE Access*, vol. 2, pp. 1553-1561, 2014.

[203] X. Meng, J. Li, D. Zhou, and D. Yang, "5G technology requirements and related test environments for evaluation," *China Commun.*, vol. 13, no. Supplement 2, pp. 42-51, 2016.

[204] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen, and P. Mogensen, "Signal quality outage analysis for ultra-reliable communications in cellular networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1-6.

[205] O. N. C. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, 2015, pp. 1190-1195.

[206] P. Guan, X. Zhang, G. Ren, and T. Tian, "Ultra-low latency for 5G - A lab trial," in arXiv:1610.04362., 2016, pp. 1-5.

[207] R. P. Liu, G. J. Sutton, I. B. Collings, "A new queueing model for QoS analysis of IEEE 802.11 DCF with finite buffer and load," *IEEE Trans. Wireless Commun.*, vol. 9, no. 8, pp. 2664-2675, Aug. 2010.

[208] G. J. Sutton, R. P. Liu, I. B. Collings, "Modelling IEEE 802.11 DCF heterogeneous networks with Rayleigh fading and capture," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3336-3348, 2013.

[209] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke, "Analysis of IEEE 802.11e for QoS support in wireless LANs," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 40-50, 2003.

[210] Z. N. Kong, D. H. Tsang, B. Bensaou, and D. Gao, "Performance analysis of IEEE 802.11e contention-based channel access," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 10, pp. 2095-2106, 2004.

[211] I. Inan, F. Keceli, and E. Ayanoglu, 2007, "Modeling the 802.11e enhanced distributed channel access function," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Nov. 2007, pp. 2546-2551.

[212] I. Tinnirello, and G. Bianchi, "Rethinking the IEEE 802.11e EDCA performance modeling methodology," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 540-553, 2010.

[213] B. Chen, J. Chen, Y. Gao, and J. Zhang, "Coexistence of LTE-LAA and Wi-Fi on 5 GHz with corresponding deployment scenarios: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 7-32, First Quart. 2017.

[214] J. Jeon, Q. C. Li, H. Niu, A. Papathanassiou, and G. Wu, "LTE in the unlicensed spectrum: A novel coexistence analysis with WLAN systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 3459-3464.

[215] Y. Li, F., Baccelli, J. G. Andrews, T. D. Novlan, and J. C. Zhang, "Modeling and analyzing the coexistence of Wi-Fi and LTE in unlicensed spectrum," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6310-6326, 2016.

[216] S. Sagari, S. Baysting, D. Saha, I. Seskar, W. Trappe, and D. Raychaudhuri, "Coordinated dynamic spectrum management of LTE-U and Wi-Fi networks," in *Proc. IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Sep. 2015, pp. 209-220.

[217] Q. Chen, G, Yu, and Z. Ding, "Optimizing unlicensed spectrum sharing for LTE-U and WiFi network coexistence," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2562-2574, 2016.

[218] A. Banchs, P. Serrano, and A. Azcorra, "End-to-end delay analysis and admission control in 802.11 DCF WLANs," Computer Communications, vol. 29, no. 7, pp.842-854, 2006.

[219] C. Cano and D. J. Leith, "Unlicensed LTE/WiFi coexistence: Is LBT inherently fairer than CSAT?," in *IEEE International Conf. on Commun. (ICC)*, 2016, pp. 1-7.

[220] F. Liu, E. Bala, E. Erkip, and R. Yang, "A framework for femtocells to access both licensed and unlicensed bands," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2011, pp. 407-411.

[221] S. Han, Y. C. Liang, Q. Chen, and B. H. Soong, "Licensed-assisted access for LTE in unlicensed spectrum: A MAC protocol design," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2550-2561, 2016.

[222] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1-5.

[223] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "LBT-based adaptive channel access for LTE-U systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6585-6597, 2016.

[224] Y. Song, K. W. Sung, and Y. Han, "Coexistence of Wi-Fi and cellular with listen-before-talk in unlicensed spectrum," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 161-164, 2016.

[225] F. Hao, C. Yongyu, H. Li, J, Zhang, and W. Quan, "Contention window size adaptation algorithm for LAA-LTE in unlicensed band," in *Proc. International Symposium on Wireless Commun. Systems (ISWCS)*, Sep. 2016, pp. 476-480.

[226] Y. Gao, X., Chu, and J. Zhang, "Performance analysis of LAA and WiFi coexistence in unlicensed spectrum based on Markov chain," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1-6.

[227] Q. Cui, Y. Gu, W. Ni, and R. P. Liu, "Effective capacity of licensed-assisted access in unlicensed spectrum for 5G: From theory to application," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, p. 1754-1767, 2017.

[228] A. M. Cavalcante, E. Almeida, R. D. Vieira, F. Chaves, R. C. Paiva, F. Abinader, S. Choudhury, E. Tuomaala, and K. Doppler, "Performance evaluation of LTE and Wi-Fi coexistence in unlicensed bands," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1-6.

[229] M. Beluri, E. Bala, Y. Dai, R. Di Girolamo, M. Freda, J. L. Gauvreau, S. Laughlin, D. Purkayastha, and A. Touag, "Mechanisms for LTE coexistence in TV white space," in *Proc. IEEE International Symposium on Dynamic Spectrum Access Networks* , Oct. 2012, pp. 317-326.

[230] B. Jia, and M. Tao, "A channel sensing based design for LTE in unlicensed bands," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2332-2337.

[231] J. Jeon, H. Niu, Q. Li, A. Papathanassiou, and G. Wu, "LTE with listen-before-talk in unlicensed spectrum," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2320-2324.

[232] G. J. Sutton et al., "Enabling ultra-reliable and low latency communications through unlicensed spectrum," *IEEE Netw.*, vol. 32, no. 2, pp. 70-77, Mar. 2018.

[233] P. Rost et al., "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72-79, 2017.

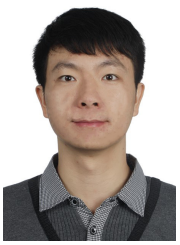
[234] J. Zeng, X. Su, J. Gong, L. Rong, and J. Wang, "5G virtualized radio access network approach based on NO stack framework," in *IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1-5.

[235] J. Zeng, X. Su, J. Gong, L. Rong, and J. Wang, "A 5G virtualized RAN based on NO stack," *China Commun.*, vol. 14, no. 6, pp. 199-208, 2017.

[236] X. Su, J. Zeng, Y. Chen, C. Gu, and L. Rong, "Evaluation of signaling loads in NO stack 5G mobile network," *China Commun.*, vol. 14, no. 7, pp. 57-66, 2017.
[237] R. D. Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102-112, Nov. 2014.
[238] T. Levanen, J. Pirskanen, T. Koskela, J. Talvitie, and M. Valkama, "Low latency radio interface for 5G flexible TDD local area communications," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2014, pp. 7-13.
[239] T. A. Levanen, J. Pirskanen, T. Koskela, J. Talvitie, and M. Valkama, "Radio interface evolution towards 5G and enhanced local area communications," *IEEE Access*, vol. 2, pp. 1005-1029, 2014.
[240] T. Shu, and M. Krunz, "Throughput-efficient sequential channel sensing and probing in cognitive radio networks under sensing errors," in *Proc. of ACM MobiCom*, Sep. 2009, pp. 37-48.
[241] H. Lee, S. Vahid, and K. Moessner, "A survey of radio resource management for spectrum aggregation in LTE-advanced," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 745-760, Second Quart. 2014.
[242] Z. Khan, J. Lehtomaki, S. Scott, Z. Han, M. Krunz, and A. Marshall, "Distributed and coordinated spectrum access methods for heterogeneous channel bonding," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 3, pp. 267 - 281, May 2017.
[243] O. Sallent, J. Pérez-Romero, R. Ferrús, and R. Agustí, "Learning-based coexistence for LTE operation in unlicensed bands," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2307-2313.
[244] N. Morozs, T. Clarke and D. Grace, "Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems," *IEEE Trans. Mobile Comput.*, vol. 15, no. 4, pp. 817-825, Apr. 2016.

Gordon J. Sutton received a B.Sc. degree in mathematics in 1993, B.E. degree in systems engineering in 1995 and PhD in control theory in 1999, all from The Australian National University.

He subsequently worked at ADC Australia designing optic fibre connectors and then at the Time Series Analysis Section of the Australian Bureau of Statistics. In 2006, he joined the Australian CSIRO as a statistician in the Quantitative Risk Management Stream of CSIRO Mathematics, Informatics and Statistics. In 2011, he joined the School of Chemistry, University of New South Wales, working in Bayesian statistics and chemometrics. Since 2015, Dr Sutton has been at the Global Big Data Technologies Centre, University of Technology Sydney, working on modelling Wi-Fi/LTE coexistence protocols. His interests include communications protocol modelling, WLAN, IoT, VANET, LTE, 5G, Markov processes, process analysis and control, forecasting, signal processing, particle filters, state space modelling and Bayesian statistics.



Jie Zeng (M'09–SM'16) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University in 2006 and 2009, respectively.

He is currently a senior engineer at the Research Institute of Information Technology, Tsinghua University. His research interests include 5G, IoT, URLLC, novel multiple access, and novel network architecture. He has authored 3 books related to 5G, and published over 100 journal and conference papers. He participated in drafting 1 national standard and 1 communication industry standard in China. He holds more than 30 Chinese and 10 international patents.

He obtained the science and technology award of Beijing in 2015 and the best cooperation award of Samsung Electronics in 2016.

Ren Ping Liu (M'09–SM'14) received his B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, China, and the Ph.D. degree from the University of Newcastle, Australia.

He is currently a Professor and Head of Discipline of Network & Cybersecurity at University of Technology Sydney. Professor Liu is also the co-founder and CTO of Ultimo Digital Technologies Pty Ltd, developing IoT and Blockchain. Prior to that he was a Principal Scientist and Research Leader at CSIRO, where he led wireless networking research activities. He specialises in system design and modelling and has delivered networking solutions to a number of government agencies and industry customers. His research interests include wireless networking, Cybersecurity, and Blockchain.

Professor Liu was the founding chair of IEEE NSW VTS Chapter and a Senior Member of IEEE. He served as Technical Program Committee chairs and Organising Committee chairs in a number of IEEE Conferences. Prof Liu was the winner of Australian Engineering Innovation Award and CSIRO Chairman medal. He has over 150 research publications, and has supervised over 30 PhD students.

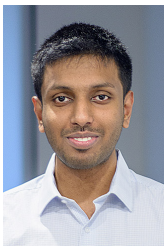
Wei Ni (M'09-SM'15) received the B.E. and Ph.D. degrees in Electronic Engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively.

Currently he is a Team Leader at CSIRO, Sydney, Australia, and an adjunct professor at the University of Technology Sydney (UTS). He also holds honorary positions at the University of New South Wales (UNSW) and Macquarie University (MQ). Prior to this, he was a postdoctoral research fellow at Shanghai Jiaotong University from 2005 – 2008; Deputy Project Manager at the Bell Labs R&I Center, Alcatel/Alcatel-Lucent from 2005 – 2008; and Senior Researcher at Devices R&D, Nokia from 2008 – 2009. His research interests include stochastic optimization, game theory, graph theory, as well as their applications to network and security.

Dr Ni has been serving as Vice Chair of IEEE NSW VTS Chapter and Editor of IEEE Transactions on Wireless Communications since 2018, secretary of IEEE NSW VTS Chapter from 2015 - 2018, Track Chair for VTC-Spring 2017, Track Co-chair for IEEE VTC-Spring 2016, and Publication Chair for BodyNet 2015. He also served as Student Travel Grant Chair for WPMC 2014, a Program Committee Member of CHINACOM 2014, a TPC member of IEEE ICC'14, ICCC'15, EICE'14, and WCNC'10.

Diep N. Nguyen is a faculty member of the Faculty of Engineering and Information Technology, University of Technology Sydney (UTS). He received M.E. and Ph.D. in Electrical and Computer Engineering from the University of California San Diego (UCSD) and The University of Arizona (UA), respectively. Before joining UTS, he was a DECRA Research Fellow at Macquarie University, a member of technical staff at Broadcom (California), ARCON Corporation (Boston), consulting the Federal Administration of Aviation on turning detection of UAVs and aircraft, US Air Force Research Lab on anti-jamming. He has received several awards from LG Electronics, University of California, San Diego, The University of Arizona, US National Science Foundation, Australian Research Council. His recent research interests are in the areas of computer networking, wireless communications, and machine learning application, with emphasis on systems' performance and security/privacy.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2897800, IEEE Communications Surveys & Tutorials



Beeshanga A. Jayawickrama received the B.E. degree in telecommunications engineering (Hons. I) and the Ph.D. degree in electronic engineering from Macquarie University, Sydney, Australia, in 2011 and 2015, respectively. He is currently a lecturer with the School of Electrical and Data Engineering, University of Technology Sydney. He has worked extensively on spectrum sensing and interference mitigation in spectrum access systems. His research interests include resource allocation, cognitive radio, and signal processing.

Xiaojing Huang (M'99-SM'11) received the B.Eng., M.Eng., and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1983, 1986, and 1989, respectively.

He was a Principal Research Engineer with the Motorola Australian Research Center, Botany, NSW, Australia, from 1998 to 2003, and an Associate professor with the University of Wollongong, Wollongong, NSW, Australia, from 2004 to 2008. He had been a Principal Research Scientist with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, NSW, Australia, and the Project Leader of the CSIRO Microwave and mm-Wave Backhaul projects since 2009. He is currently a Professor of Information and Communications Technology with the School of Electrical and Data Engineering and the Program Leader for Mobile Sensing and Communications with the Global Big Data Technologies Center, University of Technology Sydney (UTS), Sydney, NSW, Australia.

With over 30 years of combined industrial, academic, and scientific research experience, he has authored over 300 book chapters, refereed journal and conference papers, major commercial research reports, and filed 31 patents. Prof. Huang was a recipient of the CSIRO Chairman's Medal and the Australian Engineering Innovation Award in 2012 for exceptional research achievements in multigigabit wireless communications.

Mehran Abolhasan (SM) received his B.E in Computer Engineering and PhD in Telecommunications on 1999 and 2003 respectively at the University of Wollongong.

He is currently an Associate Professor and Deputy Head of School at School of Electrical and Data Engineering at UTS. He has authored over 120 international publications and has won over 3 million dollars in research funding. His Current research Interests are in Software Defined Networking, IoT, Wireless Mesh, Wireless Body Area Networks, Cooperative Networks, 5G Networks and Beyond and Sensor networks. He is currently a Senior Member of IEEE.

Zhang Zhang received the B.Eng. and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2007 and 2012, respectively. From May 2009 to June 2012, he also served as a Research Assistant for the Wireless and Mobile Communications Technology R&D Center, Tsinghua University, Beijing, China. He is currently with the Department of radio access network (RAN) research, Huawei, Shanghai, China. His current research interests include information theory and radio access technologies.

Eryk Dutkiewicz (SM) received his B.E. degree in Electrical and Electronic Engineering from the University of Adelaide in 1988, his M.Sc. degree in Applied Mathematics from the University of Adelaide in 1992 and his PhD in Telecommunications from the University of Wollongong in 1996. His industry experience includes management of the Wireless Research Laboratory at Motorola in early 2000's. Prof. Dutkiewicz is currently the Head of School of Electrical and Data Engineering at the University of Technology Sydney, Australia. He is a Senior Member of IEEE. He also holds a professorial appointment at Hokkaido University in Japan. His current research interests cover 5G and IoT networks.

Tiejun Lv (M'08–SM'12) received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1997 and 2000, respectively. From January 2001 to January 2003, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. In 2005, he was promoted to a Full Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT). From September 2008 to March 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He is the author of more than 60 published IEEE journal papers and 170 conference papers on the physical layer of wireless mobile communications. His current research interests include signal processing, communications theory and networking. He was the recipient of the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006. He received the Nature Science Award in the Ministry of Education of China for the hierarchical cooperative communication theory and technologies in 2015.