High-reliability and Low-latency Wireless Communication for Internet of Things: Challenges, Fundamentals and Enabling Technologies

Zheng Ma, *Member, IEEE*, Ming Xiao, *Senior Member, IEEE*, Yue Xiao, *Member, IEEE*, Zhibo Pang, *Senior Member, IEEE*, H. Vincent Poor, *Fellow, IEEE* and Branka Vucetic, *Fellow, IEEE*

***Abstract*—As one of the key enabling technologies of emerging smart societies and industries (i.e., industry 4.0), the Internet of Things (IoT) has evolved significantly in both technologies and applications. It is estimated that more than 25 billion devices will be connected by wireless IoT networks by 2020. In addition to ubiquitous connectivity, many envisioned applications of IoT, such as industrial automation, vehicle-to-everything (V2X) networks, smart grids and remote surgery, will have stringent transmission latency and reliability requirements, which may not be supported by existing systems. Thus, there is an urgent need for rethinking the entire communication protocol stack for wireless IoT networks. In this tutorial paper, we review the various application scenarios, fundamental performance limits, and potential technical solutions for high-reliability and low-latency (HRLL) wireless IoT networks. We discuss physical, MAC and network layers of wireless IoT networks, which all have significant impacts on latency and reliability. For the physical layer, we discuss the fundamental information-theoretic limits for HRLL communications, and then we also introduce a frame structure and preamble design for HRLL communications. Then practical channel codes with finite block length are reviewed. For the MAC layer, we first discuss optimized spectrum and power resource management schemes and then recently proposed grant-free schemes are discussed. For the network layer, we discuss the optimized network structure (traffic dispersion and network densification), the optimal traffic allocation schemes and network coding schemes to minimize latency.**

3GPP	The 3rd Generation Partnership Project
4/5G	4th/5th Generation
AWGN	Additive White Gaussian Noise
B4G/5G	Beyond 4G/5G
BA	Building Automation
BATS	Batched Sparse
BER	Bit Error Rate
BP	Belief Propagation
BS	Base Station
CA	Cooperative Awareness
CAD	Cooperative Automated Driving
CC	Convolutional Coding
CDMA	Code Division Multiple Access
CM	Cooperative Maneuver
CRC	Cyclic Redundancy Check
CS	Cooperative Sensing
CSI	Channel State Information
CSMA	Carrier Sense Multiple Access
DASE	Delay-Sensitive Area Spectral Efficiency
D2D	Device-to-Device
DC	Direct Current
Dpre	Predictive Pre-Allocation
DR	Demand Response
DSRC	Dedicated Short Range Communications
E2E	End-to-End
EDCA	Enhanced Distributed Channel Access
eMBB	Enhanced Mobile Broadband
EMS	Energy Management System
eNB	Evolved Node B
ETSI	European Telecommunications Standards Institute
EV	Electrical Vehicle
FA	Factory Automation
FEC	Forward Error Control
FIFO	First-In-First-Out
FMS	Flexible Manufacturing System
FR	Frequency Regulation
GBA	Grant-Based Access
GF	Galois Field
GFA	Grant-Free Access
GRCB	Gallager's Random Coding Bound
GSM	Global System for Mobile Communications
GSM-R	GSM for Railway
HART	Highway Addressable Remote Transducer
HP	High-Performance
HRLL	High-Reliability and Low-Latency
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IIoT	Industry IoT
IoT	Internet of Things
IT	Internet Technology
LDPC	Low-density Parity-Check
LT	Luby Transform
LTE	Long Term Evolution
LTE-A	LTE-Advanced
LTE-V2X	LTE for Vehicle-to-Everything
MAC	Medium Access Control
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
ML	Maximum Likelihood

Z. Ma is with Communications & Sensor Networks for Modern Transportation (CSNMT) International Cooperation Research Centre of China , Southwest Jiaotong University. His work was supported by National Natural Science Foundation of China (no. U1734209, no. U1709219, no. 61571373), Key International Cooperation Project of Sichuan Province (no. 2017H-H0002), Marie Curie Fellowship (no. 792406), and NSFC China-Swedish project (no. 6161101297). Ming Xiao is with Royal Institute of Technology, Sweden. Email: mingx@kth.se and his work was supported by xx. Yue Xiao is with University of Electronic Science and Technology of China, Email: xiaoyue@uestc.edu.cn and his work was supported by xx. Zhibo Pang is with the ABB Research, Sweden, Email: pang.zhibo@se.abb.com. H. Vincent Poor is with Princeton University, USA. Email: poor@princeton.edu. Branka Vucetic is with University of Sydney, Australia. Email: branka.vucetic@sydney.edu.au

mMTC	Massive Machine Type Communications
MRC	Maximum Ratio Combining
NB	Nonbinary
NBPB	NB protograph-based
NOMA	Non-orthogonal multiple access
OFDM	Orthogonal Frequency Division Multiplexing
OSD	Ordered Statistic Decoding
P2PET	Peer-to-Peer Energy Trading
PA	Process Automation
PDMA	pattern division multiple access
PEC	Power Electronics Control
PER	Packet Error Rate
PNO WSAN	Preferred Network Offload Wireless Sensor and Actor Network
PSA	Power Systems Automation
PV	Photovoltaic
QoS	Quality of Service
RB	Resource Block
RLNC	Random Linear Network Codes
RRC	Radio Resource Control
RS	Reed-Solomon
RSU	Roadside Unit
SA	Several Access
SC	Successive Cancellation
SG	Schedule Grant
SIMO	Single-Input Multiple-Output
SINR	Signal to Interference Plus Noise Ratio
SNR	Signal to Noise Ratio
SR	Schedule Request
TD	Teleoperated Driving
TD-SCDMA	Time Division-Synchronous Code Division Multiple Access
TE	Traffic Efficiency
TTI	Transmission Time Interval
UE	User Equipment
URLLC	Ultra-Reliable Low-Latency Communication
V2I	Vehicle-to-Infrastructure
V2N	Vehicle-to-Network
V2P	Vehicle-to-Pedestrian
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VE	Vehicle Equipment
VRU	Vulnerable Road User
WAVE	Wireless Access in Vehicular Environment
WCDMA	Wideband Code Division multiple Access
WIA-FA	Wireless Networks for Industrial Automation for Factory Automation
WIA-PA	Wireless Networks for Industrial Automation for Process Automation
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WirelessHP	High-Performance Wireless
WISA	Wireless Interface for Sensors and Actuators

I. Introduction

With the rapid development of computing and communication technologies, our societies and industries have become increasingly intelligent, namely, smart society or industry 4.0 (also termed smart factory). Among various enabling technologies, IoT (Internet of Things) is critical for connecting various heterogeneous devices of smart society/factory. Unlike the most of existed mobile networks designed for human-oriental communications, IoT seeks to connect large numbers of devices without or with little human intervention. The applications of IoT networks include control, intelligent identifying, locating, tracking and monitoring etc. For the heterogeneity of various applications and devices in IoT networks, the technical requirements for IoT networks are various and sometime may be rather challenging. Many application scenarios of IoT networks may require high reliability and low-latency (HRLL), such as industrial automation, vehicle-to-everything (V2X) networks, smart grids, and remote surgery, etc. In the existed systems, the devices were often connected via small-scale networks, such as Wireless Highway Addressable Remote Transducer (WirelessHART) [1], Wireless Interface for Sensors and Actuators (WISA) [2], and Wireless Networks for Industrial Automation for Process Automation (WIA-PA) [3], which are based on the IEEE 802.15.4 standard, and the WIA-FA [4], which is based on the IEEE 802.11 series standards [5]. However, these standards cannot gradually satisfy the emerging applications in terms of reliability and latency. Especially many scenarios require IoT networks to simultaneously support high reliability, low latency and massive connectivity (large scale). To meet the requirements, there have been a lot of research efforts on HRLL IoT recently.

A. Motivation and Contributions

To enable HRLL communications in IoT networks, there are roughly two evolution paths: One is based on the technical standard of public mobile cellular networks, which seeks to enhance reliability and delay performance to meet the requirements of various IoT applications with shared infrastructure. Another is for the critical applications with dedicated networks. Typical examples of the latter include high-performance wireless or Wireless HP [6], [7]. Cellular networks, which are originally designed for human-to-human communications, have evolved to meet the HRLL communication requirements in some degree. For example, ultra-reliable low-latency communication (URLLC) is one of the most innovative technical schemes for the coming 5th generation (5G) mobile network, which is quite different from its predecessors. Actually, URLLC, enhanced mobile broadband (eMBB) and massive machine type communications (mMTC) are three key use cases for 5G mobile networks [8]. Thus, URLLC can be considered as a type of HRLL communications in mobile networks. The URLLC in 5G defines the performance of a packet error rate (PER) around 10^{-6} and the end-to-end transmission delay as low as 1 ms, which could satisfy many of IoT applications with high reliability and low delay requirements. However, for many critical scenarios (e.g., industrial control and manufacturing), mobile networks may not be optimal in terms

of technologies and management. For example, the wireless transmission for industrial should potentially guarantee the PER around 10^{-9} within the transmission delay constraint as low as $10\mu s$ in the some extreme scenarios, which may be difficult for the current 5G URLLC standard. In terms of management, the high demand of responsibility for some critical applications, e.g., those in factory manufacturing, excludes the shared public networks with base stations for security reason. Then, in these scenarios, wireless HP with dedicated networks may be a good choice [6].

With the development of various new applications, the HRLL IoT networks will be more and more common, and their concrete technical requirements are quite diverse. Motivated by the promising applications, significant research has been conducted over recent years to increase the reliability and simultaneously to decrease the transmission delay. The contributions of this tutorial can be summarized as follows:

- We review comprehensively key techniques for HRLL IoT. Although different scenarios may have different performance requirement and thus different system designs for HRLL, the underline techniques they adopted share many common grounds, which have been developed substantially in recent years.
- We also give the insight for these enabling key techniques from the fundamental theory perspective. Short packet length transmission is essential to achieve the low latency. We review the information theory fundaments of short packet communications. Moreover, in addition to short length coding, we also discussed fundamental techniques adopted in physical, MAC and network layers of HRLL IoT networks. For example, we discuss the design principles for determining the appropriate packet size, obtaining preamble, grant-free access and code rate tradeoffs, choosing practical channel codes, optimizing access process and network structure etc.
- We review typical application scenarios of HRLL IoT communications, e.g., industrial automation, V2X communications and smart grids. Typical performance requirements of these scenarios are given. We also review the standardization of HRLL communications.

B. Related Literatures

Though there is no prior survey on HRLL techniques for IoT (to our best knowledge), there are overall papers on URLLC/HRLL for industrial and V2X networks, or Tactile Internet. In what follows, we shall give the brief review of these papers and also discuss the difference from our paper.

As one of 5G specifications, URLLC seeks to provide critical communication services with mobile networks. There are many tutorials on 5G URLLC [10]–[14] and its evolution roadmap. However, the application of URLLC highly depends on the standardizing process of 5G.

Several surveys focus on URLLC/HRLL for industrial use [15]–[18], namely, industrial IoT (IIoT). A few surveys present the URLLC/HRLL for V2X networks [19], [20]. They gave the use cases, application scenarios and enable techniques. However, most of them are relevant to certain specific scenarios (e.g., only on industrial control or V2X networks) and are not general HRLL IoT techniques. In contrast, this article seeks to survey the technologies common for various HRLL IoT applications.

The Tactile Internet has been described as the communication networks combining high availability, high reliability, high level of security with low latency and very short transmitting time for real-time interactive services, which is believed to be an important application of IoT. Several surveys have reviewed Tactile Internet [21], [22]. Particularly, [21] surveys the applications, requirements and challenges of industrial Tactile Internet. Tactile Internet covers both wired and wireless networks [21]. [22] considers 5G as possible URLLC solution for Tactile Internet. These surveys highlighted the state-of-art research directions and research challenges from the system and standardization aspects. However, they did not provide the design constraints and detailed underlying techniques for HRLL communications. Our survey will provide more details on the enabling key techniques.

C. Standardization of HRLL

One category of HRLL standards is based on cellular mobile networks. Actually, the concept of URLLC has not been established formally before the 5G, although a certain modification versions of cellular mobile networks have HRLL features before the 5G mobile. For instance, Global System for Mobile Communications for railway (GSM-R), which was deployed to transmit dispatch and train control commands might be the first version of such attempts. It can offer very limited data transmission capacity with 500ms end-to-end delay. Besides, for V2X networks, some of non-safety-critical services have been used in WCDMA systems by defining protocol states for Radio Resource Control (RRC) in a Vehicle Equipment (VE), but the capacity of the WCDMA system is limited, in which many VEs cannot always remain connected. LTE-V2X networks are able to support up to more than 1000 vehicles per cell in rural environments with an uplink latency below 55 ms. It also can provide a robust mechanism for mobility management. It can support a data rate of 10 Mbps with a speed upto 140 km per hour [23]. Thus, LTE-V2X can be particularly helpful at intersections by enabling a reliable exchange of cross-traffic assistance applications [24]. In recent release 14 and 15 by 3GPP, the URLLC specification is one of the key techniques. It is expected that the first 5G standard with URLLC specification will be approved by ITU in 2019 and play an important role for standardization of HRLL.

Another category of HRLL standardization is conducted by IEEE and International Electrotechnical Commission (IEC). Based on IEEE 802.15.4, the version 7 of the HART protocol (WirelessHART) is decided as wireless access for process monitoring and control applications in the industrial environments [1], [25]. WIA-PA is the Chinese industrial wireless communication standard for process automation, which was approved by IEC in 2008 and became the second wireless communication standard for the industrial in the world after WirelessHART. WISA was developed by ABB corporation and used widely in industrial field to connect devices in several

different environments [2]. It can offer less than 20ms end-to-end latency. Based on IEEE 802.11, WIA-FA was the first wireless technology specification developed specifically for factory, high-speed, automatic, control applications and officially approved in 2017 [4]. The new dedicated industrial wireless communication system, i.e. Wireless HP, is expected in the near future [6]. For V2X networks, in order to support Vehicle-to-Vehicle (V2V) connection, Dedicated Short Range Communications (DSRC), which is based on the IEEE 802.11p/1609, is used as wireless access in vehicular environment (WAVE) protocols. DSRC uses half-clocked mode with the 10 MHz bandwidth in physical layer, and borrowed Enhanced Distributed Channel Access (EDCA) idea from IEEE 802.11e to satisfy the rigorous QoS requirements in MAC layer.

The rest of the paper is organized as follows. The introductions of various applications are provided in Section II. The physical layer design for the HRLL IoT are discussed in Section III to highlight the fundamental limits and potential techniques on the physical layer. Then in Section IV, we will focus on MAC layer, where the access and transmission protocols are discussed. The analysis and design schemes for network layer will be provided in Section V. Finally, concluding remarks are given in Section VI.

II. Application Scenarios of HRLL IoT Networks

In what follows, we will discuss a few typical application scenarios for HRLL IoT networks, including factory automation, vehicular networks and smart grids in detail.

A. Industrial Automation

With the development of computing, control and communication technologies, a new generation of industry revolution, namely, industry 4.0 has largely changed our industry producing process. As a key enabling technology of industry 4.0, industry IoT (IIoT) has attracted a lot of research interests recently. Many of existing communication networks for the industries are actually based on wired networks, e.g., Ethernet or optical fiber. However, recently, a new trend of IIoT is to replace wired networks with wireless ones [6]. Motivated by promising benefits in e.g., low cost, flexibility and suitability in harsh environments or mobile scenarios, the first industrial wireless network has been implemented in real-time control applications [26]. After that, there have been lots of efforts for development and standardization for connecting devices in the industrial, e.g., WirelessHART, WIA-PA, and the WIA-FA, etc. Relative to wired networks, the advantages of using wireless ones are multi-folded: (1) Wireless networks may lead to significantly reduced costs of materials, installation, commission and maintenance; (2) Despite of channel fading or interference, wireless may be more reliable in many scenarios, e.g., the scenarios of cables subjective to aging and breaking, and easier to get redundancy links with wireless networks; (3) Wireless networks may be deployed in many scenarios where installing cables is impractical, such as moving robots, harsh industrial environments (high temperature or high voltage) and long distance (e.g., very high tower).

In emerging smart factories, IIoT is widely used to sense various environmental information and the sensed information is sent back to the controller for making decision. Then the decision based on the collected information is sent to the actuators. For many (if not the most) of these applications, latency and reliability are among the most important technical requirements. For instance, in the mining sector, remote blasting and rock-breaking control procedures are increasingly used to enhance performance and the safety of workers. Clearly, sensing and control of blasting time and magnitude are critical for efficiency and safety, which must be sensed, transmitted and processed timely and reliably. Factorial robotics are also among typical scenarios with stringent requirements on latency and reliability. Flexible manufacturing systems (FMSs) automatically adapt and react to changes in the environment, production flow, and products types. FMSs will rely on the cooperation among intelligent robots, often mounted over automatic guided vehicles. Fast running FMS is only possible with the supports of HRLL communications systems. Briefly, the basic requirements for industrial IoT networks include:

- Low latency: Many applications have rigorous demands on latency, in which short packet, simple transmitters/receivers and access protocols are preferred.
- High reliability: Some control objectives are highly valued or even dangerous, and very small transmission error could be fatal. Yet the reliability normally decreases with increasing latency requirement. For example, adopting short codeword may cause the loss of coding gain.
- Throughput: Some applications require to transmit high-resolution images or videos and thus high throughput is needed.
- Interference-robust capability: The industrial environments may be hazard. There may be strong interference generated by other communication systems and electrical equipments, e.g., powering on/off electrical engines.
- Fading-robust capability: Factory building and facilities (e.g., robot arms in assemble lines) could frequency-selectively reflect and scatter the wireless signal. This will degrade the reliability.
- Energy efficiency: Due to the low spectral density power and some terminals are power limited (power supplied by battery), energy efficiency may be critical for some applications.
- Communication range: Most of one-hop transmissions occur within 100 meters [6], [7]. Yet, some applications may need up to 1000 meters (e.g., power system protection), which may be challenging for HRLL IIoT networks.

Moreover, in many IIoT networks, the limited mobility support is acceptable. Thus the networks can be deployed statically and the channel is near-static. Other non-typical issues such as life cycle, volume, cost, heterogeneous networks configuration, security and safety should be taken into consideration as well [6].

As one of most important functionalities, IIoT is widely used for control loops in industry automation. A typical configuration of control loops is represented by a centralized wireless control network, where periodic messages are sent

from the controller to the actuators (and the sensor to the controller) with extremely low latency and high reliability. Although some general criteria can be described, a precise definition of requirements for IIoT networks has not been identified, since there are too many scenarios (and each of them may have quite different requirements accordingly). Roughly, the industrial application scenarios can be categorized as Factory Automation (FA) [27], Building Automation (BA) [28] and Process Automation (PA) [29], Power Systems Automation (PSA) [30], and Power Electronics Control (PEC) etc. BA involves all the control operations applied within buildings, such as safety, fire control, lighting, heating, water supply, air-conditioner, surveillance, energy management and property management etc. PA is quite common in chemical, pharmaceuticals, mining, oil and gas, logistics, metallurgic processes etc [29]. FA is a control system referring to run production line without or with little human intervention, including manufacturing, assembling/disassembling, packaging, palletizing, and the controlling the automatic production robot [27]. PSA focuses on controlling the generation, distribution, and transmission of electrical power [30]. PEC refers to the synchronized control of power electronics devices [31]. In Table I, we list some typical industrial automation applications, and their main technical parameters e.g., number of nodes, typical cycle time (i.e., the periodicity with which the controller, sensor, and actuators exchange data). Different scenarios pose distinct requirements, in which the BA has the lowest requirements but the PSA and PEC have the extremely high demands on latency and reliability.

TABLE I
TYPICAL INDUSTRIAL AUTOMATION SCENARIOS.

Scenario	No. of nodes	Typical Cycle Time	PER	System range
Building Automation	10^2-10^3	10 s	10^{-5}	10^1-10^2 m
Process Automation	10^2-10^3	100 ms	10^{-6}	10^1-10^2 m
Factory Automation	10^2-10^3	1 ms	10^{-6}	10^1-10^2 m
Power System Protection	10^1-10^2	100 us	10^{-9}	10^2-10^3 m
Power electronics control	10^2-10^3	10 us	10^{-9}	10^1-10^2 m

To meet high technical standards, many efforts have been put to develop various new technologies. In mobile industries, 3GPP defines the technologies of URLLC (ultra-reliable low-latency communications) [32] for the coming 5G mobile, which may also provide HRLL communications. In addition to mobile network standards for public use, to meet even higher technical standards, e.g., those for FA, PSA and PEC, some new industry wireless technologies have been proposed such as WirelessHP [6], which is aimed for dedicated industrial automation and has been deployed at certain industrial scenarios e.g., FA, moving robots and power PEC.

B. V2X IoT Networks for Transportation

With the development of various intelligent technologies, our society has never encountered such a big challenge for transportation systems before. The number of vehicles is increasing dramatically with the new wave of urbanization and the development of transportation capacity. Moreover, emission and energy-efficient regulations have been much more stringent than ever before. With the assistance of the latest wireless communication and IoT technology, it is optimistic to achieve the goal of increasing the transportation capability and efficiency. For V2X networks, the requirements on latency and reliability are stringent. For example, as one of the most important application scenarios for the 5G, the objective of V2X communication networks is to enable high-efficiency and accident-free cooperative automated driving, which shall use the available roadway efficiently. To achieve this objective, the communication networks should accommodate a diverse set of use cases, each with a specific set of requirements. The basic requirements for V2X communication networks include:

- Low latency: Though the latency requirement may not be as rigorous as certain extreme industrial control scenarios, it is still beyond the capacity of current mobile networks (e.g., 4G or below).
- High reliability: Transmission for vehicular control signaling may need extremely high reliability since the transmission errors may cause fatal accidents.
- Throughput: Some V2X applications, e.g. remote controlling and environment sensing of the traffic, require to transmit high-resolution images or videos. Accordingly, the requirements on throughput may be rather high.
- Interference-robust capability. There may be significant interference generated by other communication systems and automobile igniters.
- Fading-robust capability: Mountains and city buildings may frequency-selectively reflect and scatter the signal, which may degrade reliability further.
- Communication range: The distance of one-hop V2X transmissions may vary from dozens of meters to hundreds of meters.
- Mobility support: For city vehicles, the relative velocity may be larger than 28km per hour. For high speed trains, the speed could be more than 350km per hour. Thus, communication channels are fast time-varying. For these scenarios of high mobility, we need to design transmission schemes considering Doppler effect to improve reliability.

The most popular communication scenarios for V2X networks include [20]: 1) Vehicle-to-Vehicle (V2V) communications, in which information is exchanged among vehicles; 2) Vehicle-to-Infrastructure (V2I) communications, which occur between vehicles and roadside units (RSUs), traffic lights, and base stations; 3) Vehicle-to-Pedestrian (V2P) communications, in which vehicles communicate with people who are along the side of the road; and 4) Vehicle-to-Network (V2N), where the vehicles connect to an entity in the networks e.g., a backend server or a traffic information system.

Based on requirements, the main V2X use cases can be classified by a specific operations [33], i.e., Cooperative Awareness (CA), Cooperative Sensing (CS), Cooperative Maneuver (CM), Vulnerable Road User (VRU), Traffic Efficiency (TE), Teleoperated Driving (TD) and Cooperative Automated Driving (CAD). CA refers to warning and awareness of emergency action, e.g., emergency vehicles warning, emergency brake

warning [34], etc; CS helps to increase vehicles environmental perception by exchanging sensed data and information [35]; CM is responsible for the coordination of the route scheduling; VRU notifies the pedestrians, cyclists, non-motor vehicle objectives, etc; TE systems update routes and digital map dynamically to optimize the routes and speed by the traffic signaling systems [36]; TD enables to control a vehicle by a remote driver who controls the vehicle by sensed information and camera video from the vehicle; CAD supports to drive vehicles automatically without human intervening. The latency, throughput, reliability, and system range for these use cases are listed and compared in Table II. To meet these high requirements, a few trials have been set for demonstration. DSRC is used to connect vehicles. However, DSRC fails to guarantee the performance under severe frequency-selective multi-path and fast fading channels, and its Carrier Sense Multiple Access (CSMA) mechanism cannot avoid the unwanted interference from hidden nodes. Thus DSRC is not preferable except for V2V links. Cellular Networks for V2X use cases have attracted lots of attention recently. LTE will support V2X (namely LTE-V2X) from the release 14. It is estimated that only the most critical applications, for example, the requirements for automatic driving, are beyond the capability of the LTE-V2X (Table II). It is expected that the future 5G-V2X will cover all V2X communication requirements.

TABLE II
SYSTEM REQUIREMENTS FOR DIFFERENT V2X COMMUNICATION SCENARIOS [35]

Use Cases	Latency	Throughput	Reliability *	System Range
CA	100ms-1sec	5-96kbps	90 – 95%	< 500m
CS	3ms-1sec	5-25000kbps	> 95%	< 200m
CM	< 3ms-100ms	10-5000kbps	> 99%	< 500 m
VRU	100ms-1sec	5-10kbps	95%	< 200 m
TE	> 1sec	10-2000kbps	> 90%	> 500m
TD	5-20ms	> 25000kbps	> 99%	> 500m
AD	3ms	> 25000kbps	> 99.999%	200m

* The reliability is defined here as the packet reception ratio (PRR) within the latency requirement. [33]

A concrete V2X example is shown in Fig.1. Suppose a driver is going to use a vehicle, and he/she can hail a car via V2P link. The car can self-drive and the car is capable of CM to join platoon via HRLL V2V and V2I links. The car can also be remotely controlled by human or robot/artificial intelligence via V2I link, in which an HRLL mmWave link to support high-resolution video/images is required. During the self-driving or remote-driving phase, the car keeps exchanging sensed information and warning information with other vehicles via V2V and V2I links, as well as the traffic signaling with roadside traffic control units via V2I link. As a typical use case of URLLC, the V2X networks will have a prototype in B4G/5G.

C. Smart Grid

Smart grid refers to intelligently produce, transmit and consume electric with the aid of sensors, actuators, communication networks and central controllers. Smart grid is a power network enabling a variety of nodes of smart appliances, e.g.,

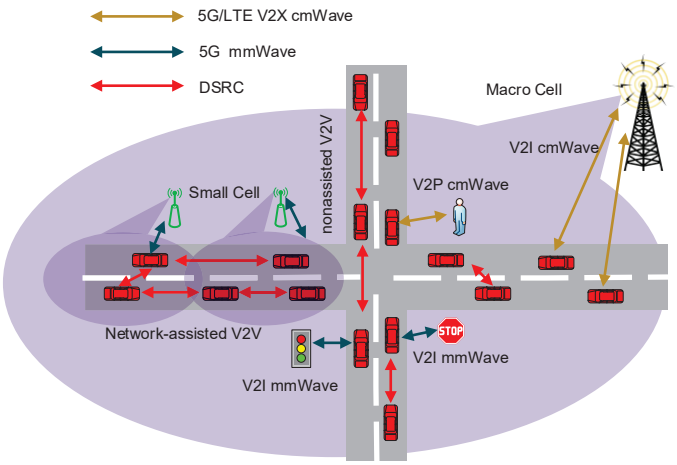

Fig. 1. A possible deployment of future V2X networks.

efficient energy generation, smart meters, smart billing and renewable energy resources. Therefore, many new applications and services are developed based on these techniques, such as Energy Management System (EMS), Demand Response (DR), Frequency Regulation (FR) and Peer-to-Peer Energy Trading (P2PET). To support two-way energy transferring in heterogeneous smart grids, the underlying communication systems should have high performance in terms of latency, rates and reliability. For instance, to enable advanced applications such as real-time pricing, a low-latency two-way real-time communication system is required. Moreover, high-rate information and energy flows due to a large number of heterogeneous prosumers (nodes able to consume and produce electrical power) for efficient distribution and storage of energy, may require response latency below 1 ms and reliability of 10^{-7} packet error rates in some scenarios [37]. To meet these requirements, there are already many related research results on the communication technologies of smart grid [38]. For instance, the most popular wired-based technologies are based on Ethernet and power line transmission while wireless technologies are based on Zigbee, WiFi, WiMAX and Bluetooth. However, the current communication technologies still cannot meet the stringent requirements of smart grid for advanced services like DR, FR and P2PET. Previously, references [39], [40] have investigated the communication system of the smart grid to meet the stringent requirement of latency. In [39], a next generation automation architecture is studied, which consists of three layers: converter level control, multi-agent system and system level layers. However, the latency and reliability of the proposed schemes are far from sufficiency. In [40], a study comparing Wi-Fi based service to wired links in a microgrid EMS was conducted. Performance analysis indicates the wireless infrastructure is more reliable, easier to build, and more scalable for small-size micro-grids, though the communication delay may be higher than the wired LAN. Reference [41] proposed IoT-grid, an architecture for a novel, programmable, small-scale Direct Current (DC) grid, which can be easily adapted in existing smart grid. The experimental platform was designed and it was observed that the processing delay of IoT devices had a

large impact on the response time of IoT-grid. Since wireless networks provide lower installation cost, faster deployment, higher mobility and flexibility than its wired counterparts, they are more favorable in most of the smart grid applications. However, there are still challenges for the design of wireless communication systems in smart grid due to heterogeneous prosumers and very high requirements in reliability, latency and rates. The main technical requirements, e.g., latency, rates, reliability, for smart grid have not been fully identified yet. It is shown recently that the existing technologies are not sufficient for a time critical smart grid with stringent demands on latency [41]. By European Telecommunications Standards Institute (ETSI), the typical response time for different smart grid functions should roughly follow the parameters in Table II-C [42]. However, there are no specifications on response time for many new scenarios of smart grid, e.g., nanogrids (smart grids in resident areas but may have complex nodes such as prosumers), which are supposed to have even higher technical standards when connected nanogrids and large grids.

TABLE III
LATENCY REQUIREMENTS FOR SMART GRID [42]

Scenario	Response time
Protection	1-10ms
Control	100ms
Monitoring	1ms
Metering/Billing	1 h-1 day
Reporting	1 day to 1 year

There are already a few survey papers on large scale smart grid, e.g., [43]. Thus, we shall concentrate on smaller scale grid, which may have complex and heterogeneous networks, namely, nanogrid as follows. A nanogrid is a very small electricity domain that is distinct from any other grids [44], which is typically serving a single building or a single load. Nanogrid could form the basis of a future electricity system built on a bottom-up, decentralized, and distributed network model rather than the top-down centralized grid we have today in most parts of the world. A central requirement of nanogrids is the ability to communicate electricity price and availability to enable matching demand with varying suppliers of electricity. Many contemporary residential buildings are integrating local energy generation, such as photovoltaic (PV), storage, electrical vehicles (EVs) and connecting to the outer power grid. Active energy consumers or prosumers, can both consume and produce electricity. Their appearance could dramatically change the future electricity system [45]. The home-based power-supplying system is the main component of nanogrid. Actually, relative to large grids, which may be quite homogeneous, nanogrid is much more heterogenous and thus has attracted lots of research efforts recently [44], [46]. Navigant Research has developed its own definition of a nanogrid as being 100 kW for grid-tied systems and 5 kW for remote systems interconnected with a utility grid [47]. A possible application case is shown in Fig.2, in which the increasing penetration of EVs has a significant impact on the electricity market. For example, vehicle batteries can serve as electricity storage nodes, supporting the stabilization

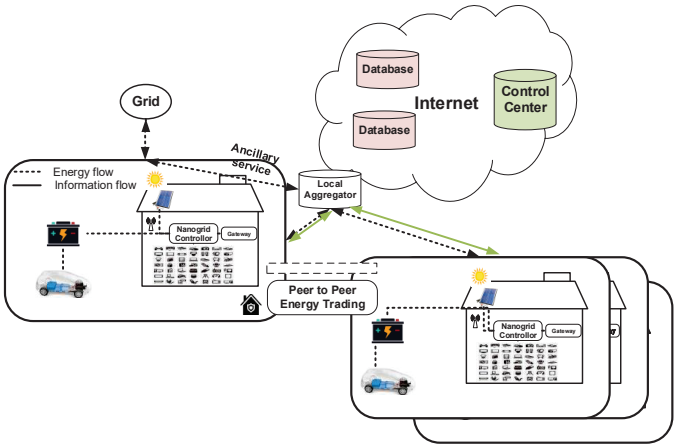

Fig. 2. An Example for Nanogrid.

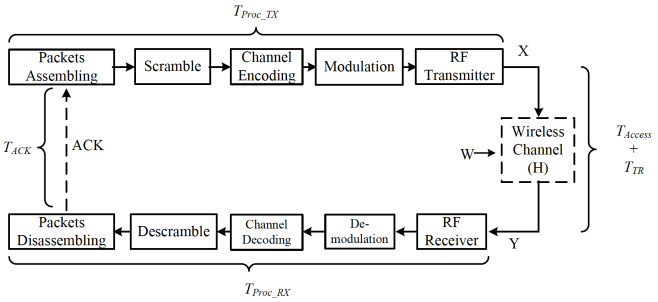

Fig. 3. System Model for HRLL Transmission in IoT.

of the electricity grid through two-way vehicle-to-grid energy transferring. These residential buildings can also become power brokers, by storing energy in battery arrays or EVs and then selling power back to the grid when it is needed. Furthermore, the nanogrid is an ecosystem integrating with the different systems, like heating, ventilating, conditioning, light control, renewable generation, storage and EVs. There can be more new services developed from nanogrid based on the communication and control systems in future. According to the ancillary services, the response time may be much shorter than 1ms for the end-to-end latency of the distribution management system of nanogrid [46].

Besides the industrial automation, V2X service and smart grid, there are also other applications which are not mentioned here for space limitation. However, above three application scenarios denote typical use cases of HRLL IoT networks, and they have shed lights on the development of future HRLL wireless technologies.

III. Physical Layer for HRLL Communications

In this section, we will first discuss the system model for IoT, and then its fundamental limits under the latency constraints. Then the more practical design principles such as frame structure, preamble design and channel coding will be given.

A. System Model

An IoT end-to-end wireless transmission system model can be shown in Fig. 3. With one/multiple antennas, the model can also be expressed as the following formula

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{W}, \tag{1}$$

where $\mathbf{X}$ denotes the transmitting complex symbols (vectors) and $\mathbf{Y}$ is the corresponding received signal, $\mathbf{H}$ is the channel gain due to fading and $\mathbf{W}$ is additive Gaussian noise. The latency defined here is the time consuming for the successful end-to-end transmission of one packet/codeword, i.e.,

$$T_{Total} = T_{Proc\_TX} + T_{Access} + T_{TR} + T_{Proc\_RX} + T_{ACK}, \tag{2}$$

where T_{Proc_TX} is the signal processing delay at the transmitter, and T_{Proc_RX} is the signal processing delay at the receiver. T_{Access} is the time to access the channel, and T_{TR} is the delay for packet transmission in the air, and T_{ACK} is the time to receive the acknowledgment. Generally speaking, T_{Proc_TX}, T_{Proc_RX}, T_{Access} and T_{TR} highly depend on packet length due to following reasons:

- Current transmitting and receiving schemes normally are packet-wise. A shorter packet leads to less processing and transmission delay, namely, smaller T_{Proc_TX}, T_{Proc_RX} and T_{TR}, and vice versa.
- Channel coding is packet-based as well, and the encoding/decoding algorithms of channel coding are very time consuming. Short packet implies short channel code schemes, which can reduce the delay.
- For modern channel coding e.g., Turbo codes and Turbo product codes, an interleaver is normally used within a packet/codeword. The delay is very sensitive to interleaver size. Short packet results in short interleaver sizes with low latency (assuming no inter-packet interleaving).

Thus for low latency communication, the length of packet should be short. For some IoT networks, T_{Total} is strictly constrained by data refresh rate. Thus, a complete packet should be transmitted/received successfully within limited time interval.

There are many factors could affect the transmission reliability in physical layer, such as modulation, interleaver, channel coding, detection schemes and decoding algorithms, etc. Some of these factors are highly related to latency. For example, the signal processing delay T_{Proc_TX} and T_{Proc_RX} highly depend on interleaver size and channel encoding/decoding algorithms. The sophisticated algorithms could enhance reliability, but may deteriorate latency and power consumption. Especially, many devices are power limited in IoT.

B. Fundamental limits for HRLL communications in IoT

In wireless communication systems, it is a challenging task to achieve high reliability and low latency simultaneously, particularly for resource limited communications e.g., IoT. Many traditional techniques (e.g., strong channel codes) have been proposed to improve reliability, but often have to sacrifice latency. On the other hand, reducing latency with short packet length could cause decreasing reliability, because short block length cannot secure the large coding gain, and the size of overhead symbols in packets (metadata), such as pilot symbols, header and preambles, may be comparable with information length.

Fundamental results in information theory [48] show that when the packet length goes to infinite, there always exist a channel coding scheme, with which the transmitting symbols can be recovered with arbitrarily small error probability, if the communications rates are equal to or smaller than channel capacity, which is defined as

$$C = \lim_{\epsilon \to 0} \lim_{n \to \infty} R^*(n, \epsilon), \tag{3}$$

where n is the code length and ϵ is the error probability and $R^*(n, \epsilon)$ is the maximal rate with n and ϵ. However, [48] does not consider the scenarios with finite packet length, namely, finite n. In [49], the concept of error exponent is proposed and error probability bounds are analyzed as the function of R^* and n. In recent years, the fundamental limits of short packet transmission have been studied and the significant progress have been made. In [50], the tight bounds for the maximal coding rate over various channels were derived for finite block length. For AWGN channels, the maximal coding rate for finite packet transmission was provided by [50]:

$$R^*(n, \epsilon) = C - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon) + O\left(\frac{\log(n)}{n}\right), \tag{4}$$

where $Q^{-1}(\cdot)$ is the inverse of Gaussian function, and the capacity C and channel dispersion V are given by the functions of SNR ρ

$$C(\rho) = \log(1 + \rho), \tag{5}$$

$$V(\rho) = \frac{2 + \rho}{(1 + \rho)^2} (\log e)^2 \rho. \tag{6}$$

Substituting (5) and (6) into (4), and using normal approximation $\frac{\log n}{2n}$ to approximate $O\left(\frac{\log(n)}{n}\right)$, we have

$$R^*(n, \epsilon) \approx \log(1 + \rho) - \sqrt{\frac{\frac{2+\rho}{(1+\rho)^2} (\log e)^2 \rho}{n}} Q^{-1}(\epsilon) + \frac{\log n}{2n}. \tag{7}$$

To better clarify how the blocklength affects the performance, the rate changes along the packets error probability ϵ and blocklength n are plotted in Fig.4 for one example.

Although the normal approximation is not applicable to determine how long a packet is needed for a certain packet error probability in accuracy, it is still good enough to show the trend. As can be seen in Fig.4 the maximum rate drops sharply when blocklength and packet error probability decreases. For example, for blocklength $n < 50$, it is very difficult to achieve $\epsilon < 10^{-9}$ in AWGN channel.

For block fading channels, the situation may get even worse. In [51], the rate for block fading channel was analyzed and it was shown that

$$R^*(n, \epsilon) = C_\epsilon + O\left(\frac{\log(n)}{n}\right), \tag{8}$$

where C_ϵ is the outage capacity, which defines the supremum of all rates R satisfying $P_{out}(R) \leq \epsilon$, i.e.

$$C_\epsilon = \sup\{R : P_{out}(R) \leq \epsilon\}, \tag{9}$$

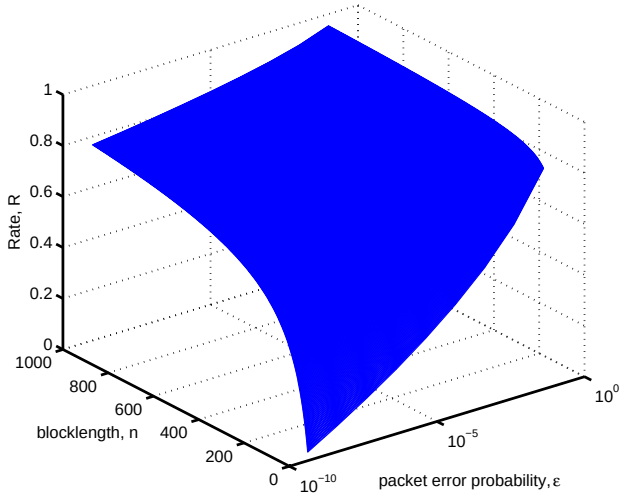

Fig. 4. Normal approximation on maximal coding rate $R^*(n,\epsilon)$ for AWGN channel by using (7) with SNR $\rho = 0$ dB.

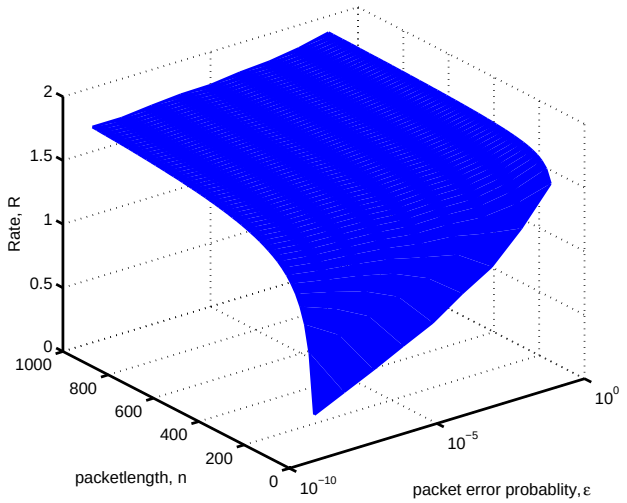

Fig. 5. Normal approximation on maximal coding rate $R^*(n,\epsilon)$ for block fading channel by using (7) with 2×2 MIMO and SNR $\rho = 10$ dB.

where $P_{out}(R) = \mathbb{P}\left[\log\left(1+|\mathrm{H}|^2\rho\right) < R\right]$ is the outage probability. A normal approximation can also be applied to get the solution of

$$\epsilon = \mathbb{E}\left[Q\left(\frac{C\left(\rho \mid H \mid^2\right) - R^*(n,\epsilon)}{\sqrt{\frac{V(\rho|H|^2)}{n}}}\right)\right]. \tag{10}$$

The approximation for block fading channels is plotted in Fig.5. We can see from Fig.5 that MIMO can help to achieve the maximum rate. Yet when the packet length is shorter than a certain value at a certain packet error probability, the rate loss falls dramatically and the effective information transmission becomes impossible.

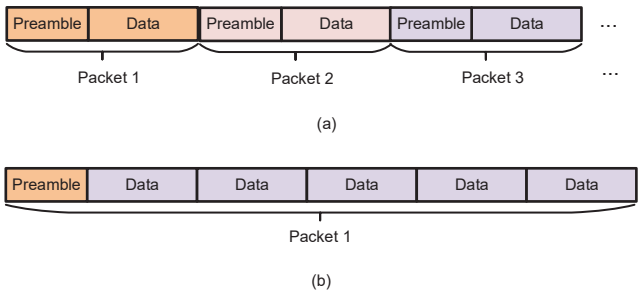

Fig. 6. Two possible frame structure approaches for downlink low-latency transmissions in IoT: (a) The message of individual device is encoded into a separate packet with individual preamble; (b) The messages of all devices are jointly encoded in a single packet but share only one preamble.

C. Frame structure and preamble design

The frame structure and preamble deign are quite challenging for low latency IoT. Many advanced receivers require accurate channel state information (CSI), which is normally obtained by channel estimation. However, low latency transmission normally cannot allow dedicated time slots for channel estimation for finite block length. To cope with this problem, in downlink transmission, joint encoding all data symbols and one copy of preamble (for estimating channels) into one packet is preferred as shown in Fig.6 [10]. However, this scheme requires all receiving nodes to decode all information symbols, which may consume lots of power and increase decode latency. Indeed, there is a trade-off between reliability and power consumption, and [52] discussed the trade-off in detailed.

In the uplink, Lee et al. [53] proposed to use highly reliable symbols as partial preambles and a powerful iterative receiver is used to enhance the transmission performance for short frames. The essential idea is how much overhead (preamble, and error control coding redundancy symbols) is optimal in a short frame. A pragmatic approach to the problem is proposed in [54]. The frame structure model is shown in Fig.7. The total frame consists of N symbols, in which k symbols are information symbols, $n-k$ symbols are redundancy symbols generated by channel coding, and L symbols are preamble used for channel estimation. From an information-theoretical point of view, longer preamble leads to more accurate CSI, which helps to increase detection performance. The longer codes will also increase the decoding reliability as well. Unfortunately, using short frame can guarantee neither long preamble nor long codes in a frame. Thus there must be a trade-off between the length of preamble and codes, i.e., optimum L for the length of preamble. The tradeoff can be presented by the so-called *mismatched decoding metric*, in which the imperfect acquisition of CSI caused by insufficient pilot symbols is considered in the decoding process. Gallager's random coding bound [49] and sphere packing bound were taken as the upper and lower bounds for mismatched decoding framework, respectively. For roughly estimating how long a preamble is necessary for a certain length packet, we employs Gallager's random coding bound as the reference. The mismatched decoding means that the general detection/decoding model (in Fig. 3) does not have to know channel perfectly. That is, the

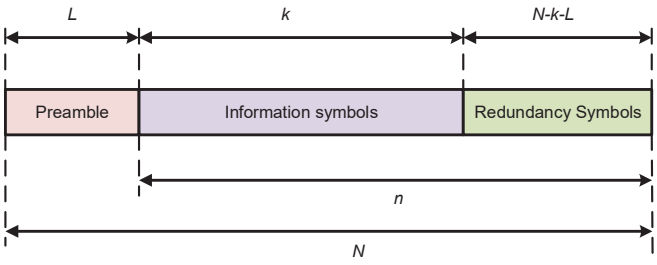

Fig. 7. A frame structure for pragmatic approach [54]. A frame consists N symbols, which are divided into L-symbol preamble, k-symbol information data and $(N-k-L)$ redundancy.

channel estimation $\hat{\mathbf{H}}$ may not match the CSI matrix $\mathbf{H}$. Let $\hat{h}$ be a realization of $\hat{\mathbf{H}}$, then channel estimation error $\Delta = h - \hat{h}$ is complex normal distributed with zero mean and variance $2\sigma^2/L$. For such a channel estimation $\hat{h}$, it is discovered from [55] that the Gallager's random coding bound is

$$P_G = 2^{-nE_G(R_c,\hat{h})}, \tag{11}$$

where

$$E_G(R_c,\hat{h}) = \max_{0\le\rho\le 1}\sup_{s\ge 0}\left(E_o(s,\rho,\hat{h}) - \rho R_c\right), \tag{12}$$

where

$$E_o(s,\rho,\hat{h}) = -\log_2 \mathbb{E}\left[\left(\mathbb{E}\left[\left(\frac{W(Y|X';\hat{h})}{W(Y|X;\hat{h})}\right)^s \middle| X,Y\right]\right)^\rho\right], \tag{13}$$

where $W(Y|X;\hat{h})$ is the complex Gaussian transition probability function under imperfect channel estimation $\hat{h}$

$$W(Y|X;\hat{h}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}|y-\hat{h}x|^2\right). \tag{14}$$

By Gallager's random coding bounds, we can find the suitable length of preamble when the packet size is fixed. For example, Fig.8 and Fig.9 show the SNR required to achieve a PER of 10^{-5} for various preamble lengths for packet lengths are 128 and 256, respectively. The channel code rate is assumed to be 0.5 and the channel is AWGN. It can be inferred from the figures that the preamble length L around 12 and 25 might be the optimum for packet lengths $N = 128$ and $N = 256$, respectively.

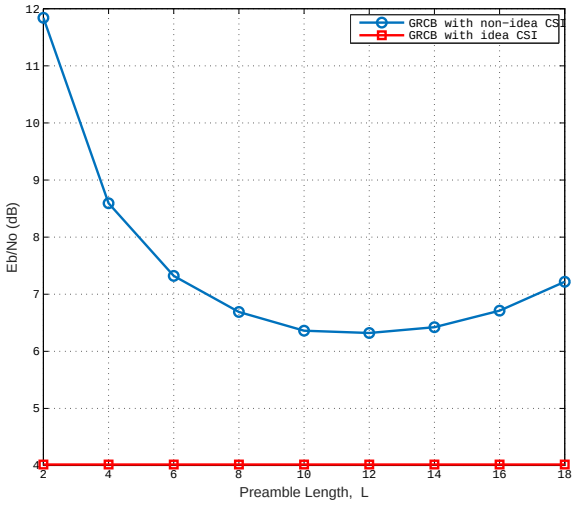

Fig. 8. SNR required to achieve a PER of 10^{-5} for various preamble lengths. Packet length $N = 128$, the code rate is 0.5 and AWGN channels.

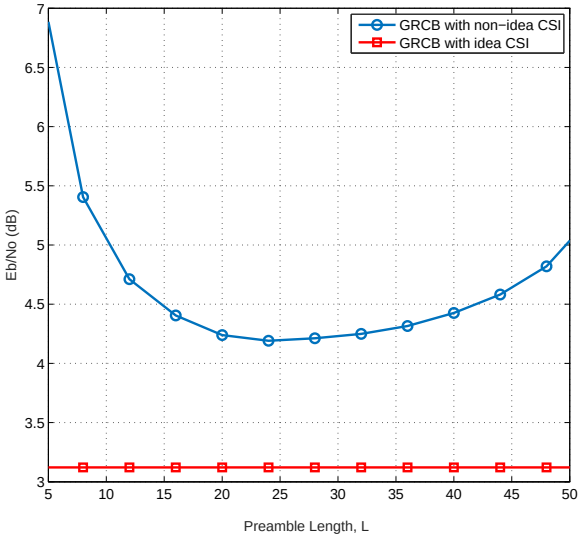

Fig. 9. SNR required to achieve a PER of 10^{-5} for various preamble lengths. Packet length $N = 256$, the code rate is 0.5 and AWGN channels.

D. Channel coding for HRLL Communications

To achieve high reliability, channel coding is a natural choice since with low-latency constraints, retransmission may not be optimal. As a forward error control (FEC) strategy, channel coding has been developed for many years, from algebra codes (e.g, BCH codes, Reed-Solomon codes), convolutional codes (CC), to more recently Turbo codes, sparse matrix-based codes (e.g., LDPC codes, Fountain codes) and polar codes. In Table. IV, we give the timeline and applications of some typical channel codes, especially those related to wireless communications. To achieve high reliability and large throughput, modern coding schemes (e.g., Turbo codes, LDPC and Polar codes) normally have large block length, namely, in terms of tens of thousands bits per codeword [56], which may not be suitable for low-latency applications. Thus, efficient channel coding schemes with finite length should be developed for HRLL communications. Moreover, the encoding and decoding latency is also important for HRLL communications. Furthermore, the energy-efficient coding schemes are also important, for certain IoT scenarios, e.g., battery-powered sensor networks. In what follows, we will first give a brief review on principles of HRLL channel codes, and then a few typical channel coding schemes will be discussed for HRLL communications.

1) Design Principles:

- With finite block length, the error rates will be non-negligible, which however should be reduced to a reasonable level. Turbo and LDPC (low-density parity-check)

TABLE IV
A BRIEF REVIEW OF CHANNEL CODES FOR VARIOUS APPLICATIONS AND TIMELINES. HERE DSC, RM, TCM STAND FOR DEEP SPACE COMMUNICATIONS, REED-MULLER, TRELLIS CODE MODULATION, RESPECTIVELY.

Year	App.	Codes	Year	App.	Codes
1950	Data Storage	Hamming Code	2008	Wireless HART	No
1970s	DSC	R-M Codes	2009	IEEE 802.11n	CC, LDPC
1980s	TCM	CC	2009	PNO/WSAN	Repeated codes CRC
1992	GSM	CC	2011	3GPP V10.1.0	CC Turbo codes
1999	IEEE 802.11a/b	CC	2012	IEEE 802.16m	CC, Duo-binary Turbo codes
2000	CDMA 2000	CC Turbo codes	2012	LTE-Advanced	CC Turbo codes
2000	TD-SCDMA	CC Turbo codes	2012	IEEE 802.11ac	CC, LDPC
2000	WCDMA	CC, RS+CC Turbo codes	2012	IEEE 802.11ad	RS, LDPC
2003	IEEE 802.11g	CC	2014	ISA100.11a	No
2006	IEEE 802.16e	Turbo codes LDPC	2016	5G draft	LDPC Polar codes

like codes are more recently developed coding schemes. However, they normally have error floors with finite block length (i.e., a few hundred bits or even shorter) at medium-to-high SNR regions. Some solutions have been proposed to mitigate the problem. One solution is to increase the coding field size, namely, from binary codes to high order Galois fields, such as nonbinary (NB) Turbo codes and NB-LDPC defined over GF(256). With larger field sizes, the minimum Hamming distance is enlarged, and thus lower error floors are achieved, compared to their binary counterparts. Another solution is concatenated or evolved coding schemes, such as concatenated Reed-Solomon (RS) and CC codes, RS-based LDPC codes, and LDPC-CC, which show improved error floor performance, relative to individual component codes.

- The decoding latency should be considered, especially from a practical implementation aspect. By increasing to higher field sizes, or the concatenation of two coding schemes, the error floor performance can be improved. However, the decoding latency may be increased. For example, RS and NB-Turbo concatenated codes have better error floor performance. However, computation complexity (and latency) may be greatly increased in high field sizes. For coding schemes with iterative decoding algorithms, the schemes with parallel processing capability, e.g., LDPC codes are preferred, relative to those with sequential processing, e.g., Turbo codes.
- The coding schemes should also be optimized for burst error correction [7]. In typical IoT application scenarios, due to the movement of the vehicles, electromagnetic radiation from switching on/off of high power appliances and coexistence of different wireless devices, there may be strong continuous interference leading to burst errors. To combat burst errors, interleaving/de-interleaving is often used. However, for finite code length, the error corrupted bits by continuous interference cannot be scattered far apart enough. Hence there may be still burst errors in the de-interleaved packets.
- For battery-powered IoT devices, the codes should be energy efficiency. If the extra power consumption for channel encoder/decoder exceeds the transmitted power savings due to using the codes, then the codes may not be energy-efficient compared with an uncoded system. Thus the strong codes with high code rate are preferred, since they can offer the better reliability with less redundancy bits. The candidates coding schemes could be LDPC or polar codes. Except for adopting low complexity decoding algorithms to reduce the power consumption at decoder, decoding control schemes, such as stopping criteria [57], [58], can be employed to reduce the unnecessary computations for decoders.

2) Advances of Codes for HRLL Communications: In what follows, we will discuss some commonly used codes for HRLL communications.

- LDPC types of codes. Actually, in terms of decoding latency, LDPC codes may be preferable for the inherent parallel decoding. In [59], finite-length NB protograph-based (NBPB) LDPC codes are proposed, which choose the edge weights and show coding gains in about 256 bits of code length. In [60], NB-LDPC codes in a field of GF(256) have a parallel decoding structure and greatly reduce decoding latency. As rateless codes, LT codes and Raptor codes [61] show flexibility in error controlling capability and are used in multimedia dissemination. Raptor codes show excellent performance for large block length with iterative decoding in erasure channels. For finite block length, Raptor codes also show good performance

with ML decoding [61]. However, for noise channels, the results for finite length Raptor codes are still limited.

As a recent breakthrough in coding theory, polar codes proposed in 2009 theoretically show the capacity-achieving capability for symmetric binary-input memoryless channels [62]. The basic decoding scheme for polar codes is the SC (successive cancellation) algorithm. However, for BER/PER performance, polar codes may not necessarily be better than Turbo codes or LDPC codes for finite packet lengths, even with high-complexity ML decoding. Thus, a cyclic redundancy check (CRC) aided decoding strategy was proposed and demonstrated a gain of about 0.5-1 dB over Turbo codes and LDPC codes, for the rate-1/2 polar code with a packet length of 1024 bits at PER 10^{-4} [63], [64]. For polar codes with finite length, the SC listing algorithm, the ordered statistic decoding (OSD) algorithm, and the SC stacking algorithm are proposed [65]–[67]. These algorithms have greatly reduced decoding complexity. As a result, the CRC-aided polar codes have been proposed in the 5G standard for short block length communications. Meanwhile, the SC decoding algorithm can operate in parallel by exploiting combinational logics, and thus shorten decoding latency [68]. For a polar code with packet length of 1024 bits, the parallel decoding structure can reach a throughput of 2.5 Gbps. To achieve higher throughputs while avoid BER/PER loss, a better alternative is the highly paralleled belief propagation (BP) decoding algorithm. Based on the sub-factor graph freezing technique proposed in [69], the average number of iterations is reduced to obtain a throughput of 13.9 Gbps, for a rate-1/2 polar code with packet length of 1024 bits.

- Classic algebraic codes

  With the advantage of their algebraic structures facilitating hardware implementation, cyclic codes have been widely applied. BCH codes are among the most important cyclic codes. With low processing latency, flexible code rates and packet lengths, BCH codes are used in flash memory and optical communications [70]. In general, for algebraic codes, decoding complexity is lower for a coding scheme defined on a smaller field size, but also has less error correction capability for burst errors. RS codes, a subclass of BCH codes defined over high order GF(2^M) with M-bits each symbol, have showed excellent burst error correction capability, and have been applied in industrial IoT for various scenarios [71]. With similar code rates, RS codes have better BER/PER performance than BCH codes, and the improvement becomes larger with increased SNR. In medium-to-high SNR regions, RS$(63, 55)$ codes with a code rate about 0.88 outperform the BCH$(63, 45)$ code with a code rate of about 0.7.
- Convolutional Codes (CC)

  In contrast to block codes (e.g., LDPC, RS codes), the encoder and decoder of CC are continuous and do not have to wait the whole codewords for delivering encoded and decoded symbols. Thus, CC has an inherent advantage over block codes in terms of latency. For short packet communications, e.g., less than 100 information bits, CC can approach the lower bounds of block codes, and show better BER/PER performance than binary block codes [72], [73]. With the larger number of states, constraint length, and decoding window length, better BER/PER performance can be achieved [74], [75] for CC. However, meanwhile, the decoding latency will increase correspondingly. Latency comparisons between CC and block codes have been investigated in [72], [76]. Although ML and iterative decoding can offer optimal or near optimum BER/PER performance, considering the tight latency requirement, Viterbi and stack sequential decoding are still the best choices in practice [77], [78]. For CC with constraint length no more than 10, Viterbi decoder is preferred, but due to the exponential increased complexity with respect to constraint length, the stack sequential decoder performs better for larger constraint length (for example, larger than 10).

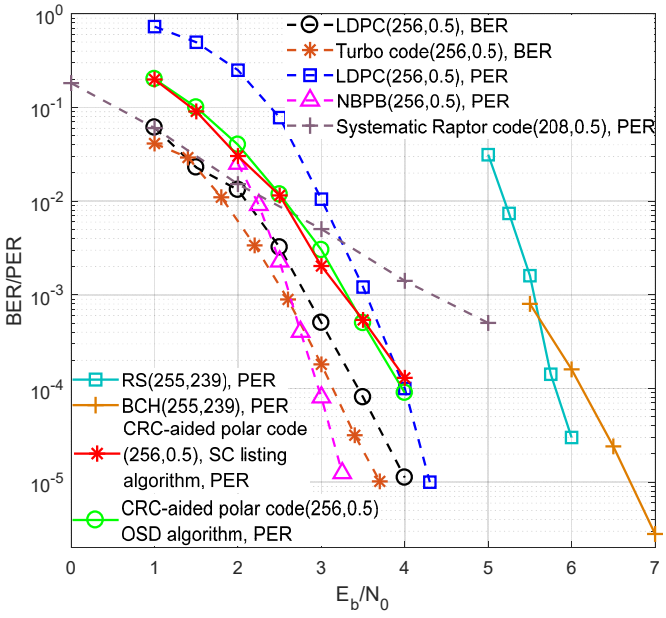

Fig. 10. Error rates (BER/FER) of LDPC-type codes, algebraic codes and CC with finite block length in AWGN channels

In Fig. 10, we compare the error rates (BER/FER) of LDPC-type codes, algebraic codes and CC with finite block length in AWGN channels. From simulation results, we can see that non-binary LDPC (e.g., NBPB) codes have excellent performance in PER and they also outperform Polar codes for finite length. Systematic Raptor codes are not suitable for finite length coding. Both Turbo codes and LDPC codes have excellent performance in term of BER.

IV. MAC Layer Design for HRLL IoT Networks

One of largest challenges in current IoT networks is to meet increasing reliability and latency requirements under frequency and energy resource constraints. Especially, the IoT networks may be heterogeneous and power limited (e.g., those powered by batteries). Therefore, in addition to improving the physical layer, we should also develop the resource-efficient

scheduling in the MAC layer for HRLL IoT networks. The resource-efficient scheduling of IoT networks mainly focuses on spectrum-efficient and energy-efficient management. These resource-efficient management techniques aim to maximize the spectral efficiency and energy efficiency, with latency and reliability constraints. In what follows, we shall review the resource-efficient management solutions for HRLL and the efficient congestion control mechanism design for reducing scheduling (queuing) latency, as well as a promising grant-free scheduling scheme in details.

A. Spectrum-efficient and Power-efficient Resource Management

Currently, the spectral efficiency is one of the main performance indicators for designing and optimizing wireless networks. In what follows, we shall provide the review of spectrum-efficient scheduling strategies for mission-critical communication and discuss their suitability for HRLL IoT communications. In [79], [80], the schemes of minimizing QoS-constrained spectrum allocations were proposed for HRLL communications. Particularly, reference [79] adopted the statistical multiplex queueing mode to reduce the bandwidth for ensuring queueing delay requirement, by which the packets to different devices were waiting in one queue at the buffer of transmitting nodes. The bandwidth scheduling was first optimized with given delay components, and then the uplink and downlink bandwidth was optimized for the given end-to-end latency. The simulation results showed that the joint two-step algorithm required half of the total bandwidth of the non-joint optimization. Based on above approaches, reference [81] applied the delay-sensitive area spectral efficiency (DASE) as the objective function, which sought to minimize the total bandwidth consumed by the devices and simultaneously ensure the strict QoS constraint on reliability such that the DASE was maximized. Moreover, to achieve the different requirements associated with different application scenarios, the HRLL communications schemes with various application scenarios such as industrial process automation [82], factory settings [83], and typical indoor environments [84] were investigated. Additionally, from the view of spectrum effectiveness, the HRLL communications related to utilizing the unlicensed spectrum were surveyed in [85].

Since energy efficiency is a very important metric in wireless IoT networks under latency and reliability constraints, and many results are reported on energy efficiency and HRLL designs for various IoT applications [86]–[88]. In [86], a scheme of joint optimized transmit power, bandwidth and the number of active antennas for energy efficiency in HRLL communications was studied. For HRLL V2V communications, a RSU-assisted virtual clustering mechanism was proposed in [87], where RSU grouped vehicles into pairs of virtual zones over the set of allocated RBs. The proposed mechanism aimed to allocate the resources of each vehicle such that the total power consumption of the system was minimized, and meanwhile queueing latency and reliability constraints are satisfied. In [89], for achieving HRLL V2X transmission,

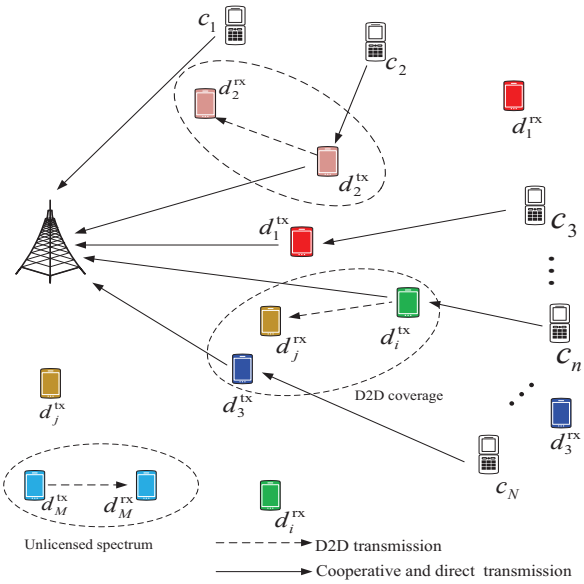

Fig. 11. An OFDMA cellular network with D2D communications. The solid line shows the cooperative transmission between CUs and D2D transmitters and direct uplink transmission for CUs while the dashed line shows the D2D pair links. The D2D users with the same color represent the initial D2D pairs.

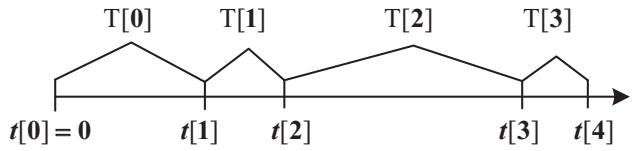

Fig. 12. Timeline illustration renewal frames for the system.

energy consumption was reduced through a semi-centralized and distributed dynamic power allocation scheme. Moreover, it was shown that there was tradeoff between latency and energy efficiency. In [90], the inherent energy-latency tradeoff was investigated in HRLL communications with retransmissions.

For the use cases we discussed before, we summarize the resource management for HRLL communication in Table.V, where the specific latency includes queuing delay, access delay, feedback delay, transmission/retransmission delay, and end-to-end (E2E) delay. However, there may be tradeoffs between latency and reliability. For instance, improving reliability by retransmissions and opportunistic radio channel aware scheduling techniques may increase latency.

To highlight the importance of spectrum-efficient and power-efficient in HRLL IoT communications, we will give an example on the dynamic uplink transmission for an OFDM access (OFDMA) cellular network with D2D communications [103]. Consider the uplink of a cellular network consisting of a single BS and a set $\mathcal{N} = \{c_1, c_2, \ldots, c_N\}$ of N cellular users (CUs) located in the edge area under its coverage, as shown in Fig. 11. There is a set $\mathcal{D}(\mathrm{tx}) = \{d_1^{\mathrm{tx}}, d_2^{\mathrm{tx}}, \ldots, d_{M_1}^{\mathrm{tx}}\}$ of M_1 D2D transmitters, and the set of M_2 D2D receivers is denoted as $\mathcal{D}(\mathrm{rx}) = \{d_1^{\mathrm{rx}}, d_2^{\mathrm{rx}}, \ldots, d_{M_2}^{\mathrm{rx}}\}$. The D2D transmitters consume transmit power to relay mobile cell-edge cellular users for uplink transmission in exchange for bandwidth from cellular users for D2D communications. In [103], the D2D pairs can

TABLE V
SUMMARY OF IMPORTANT SURVEYS ON RESOURCE MANAGEMENT FOR HRLL

Case	Reference	Approach	Main contribution
Queuing delay	[84]	Reduce inter-cell interference	Reliability can be enhanced via dealing with inter-cell interference effectively for a typical indoor environment.
	[91]	Path/route selection and rate allocation	The decrease of latency is up to 50.64% and 92.9% with a guaranteed probability of 99.9999%. *No explicit focus on IoT*.
	[92]	Packet schedulers	Applying the scheduler at an eNB each TTI, some network-wide utility function can be improved by allocating RBs to various flows/UEs.
	[93], [94]	Lyapunov concepts based dynamic algorithm	Propose utility-delay control approach resorted on Lyapunov concepts with joint variable channel condition and queue.
	[95]	MEC and caching	A comprehensive survey on joint radio-and-computational resource management for heterogeneous services. *No explicit focus on IoT*.
Access delay	[82]	Pre-allocation scheme based on the semi-persistent scheduling technique	Focusing on the uplink access in industrial process automation, the low spectrum utilization issue can be addressed by DPre.
	[96]	Enhanced Grant-free transmission	Describes various design to enhance the reliability for low latency grant-free uplink data transmission including repeated transmissions, feedback control, etc.
Retransmission, feedback delay	[89]	Retransmissions in the finite block-length regim	A discussion of energy-latency tradeoff in URLLC, along with the promising solution for it, which considers the number of rounds, the blocklength and the transmit power.
Transmission delay	[97]	Variable-rate pilot and diversity	An investigation on how to exploit variable-rate URLLC scheme for improving spectral efficiency.
E2E delay	[79]	Statistical multiplexing queuing mode	An elaboration of packet delivery mechanism from an E2E perspective on technology and promising solution. *No explicit focus on IoT*.
	[81], [86], [98]	Proactive packet dropping mechanism	URLLC-constrained transmission design by joint optimizing the packet error probability, queueing delay violation probability, and packet dropping probability.
	[99]–[102]	Diversity/short transmission intervals	Diversity via which high reliability communications in a fading channel can be enhanced by utilizing variations in time, frequency, and space to ensure communication robustness.

move around within the area covered by the BS, while the CUs move around the edge area of cellular network, which can be regarded as a renewal system when the system state is refreshed. For every frame, a new peer selection strategy is implemented that affects the frame size. Thus, the infinite-horizon optimization problem has variable length renewal frames. The time slots between two consecutive selection form a frame and the successive frames of duration $\{T[0], T[1], ...\}$ are shown in Fig. 12. Define $t[0] = 0$, and for each positive integer r defined in $t[r]$ as the time at which the selection process event is triggered for the rth time as

$$t[r] \triangleq \sum_{i=0}^{r-1} T[i]. \tag{15}$$

Under spectrum-power trading scenario above, the energy efficiency of CU is defined as the ratio of its achieved data rate and its overall power consumption, i.e.,

$$\mathrm{EE}_n^{\mathrm{CU}} = \frac{R_n^{\mathrm{CU}}}{\hat{p}_n \sum_{i=1}^{M} \left(1 - y_n x_{d_i^{\mathrm{tx}}}^{(n)} \omega_{d_i^{\mathrm{tx}}}^{(n)}\right) B_n}, \tag{16}$$

where B_n and $\hat{p}_n$ are the assigned bandwidth and the fixed transmit power density of CU c_n, respectively. $x_{d_i^{\mathrm{tx}}}^{(n)}$ is the D2D transmitter, and d_i^{tx} is the relay selection indicator of cellular user c_n, defined as

$$x_{d_i^{\mathrm{tx}}}^{(n)} = \begin{cases} 1, & \text{if } d_i^{\mathrm{tx}} \text{ is selected for relaying CU } c_n, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

$y_n \in \{0, 1\}$ denotes the cooperative D2D communication mode selection for CU $c_n \in C$. $\omega_{d_i^{\mathrm{tx}}}^{(n)}$ is the trust level between CU c_n and the D2D transmitter d_i^{tx}.

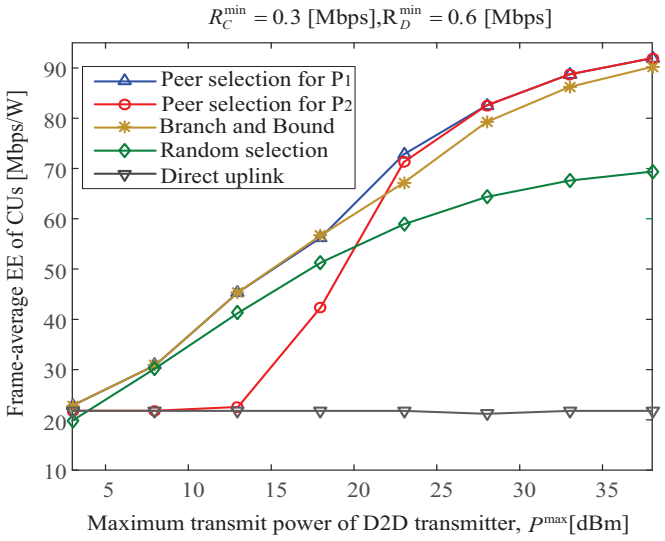

Fig. 13. Average EE of cellular users versus maximum transmit power of the D2D transmitters.

Our goal is to maximize the average energy efficiency of cellular users, while high-reliability is achieved. Fig. 13 shows the average energy efficiency of cellular users achieved by the Lyapunov based drift-plus-penalty algorithm, the branch-and-bound scheme, the random selection scheme, and the scheme that only selects the base station. From simulation results, we can see that the average energy efficiency of cellular users achieved by the drift-plus-penalty algorithm and the branch-and-bound scheme increases as the maximum transmit power, $p^{\max}$, increases, and they outperform the other two schemes.

B. Congestion Control Mechanism Design

For MAC layer, it is desired to design an efficient congestion control mechanism, to achieve both high utilization and fairness while guaranteeing low bottleneck queue length and minimizing congestion-induced packet drop rate [104]–[106]. For instance, the authors in [104] proposed an adaptive congestion control protocol by combing the estimation of the bottleneck queue size and a measure of fair sharing for high bandwidth-delay product networks. In order to achieve the minimum queue length, minimum network latency, and high link utilization, a congestion control mechanism based on heterogeneous flow with different packet sizes and different round-trip times was presented in [105]. Considering sharing the bandwidth, a price-based distributed congestion control was introduced in [106] to maintain a bounded queueing delay when competing with other delay based flows, and avoid starvation when competing with loss based flows. More and more researchers exploited the underlying support of multi-channel communication to address the delay problem [107]–[110]. Specifically, a channel access policy based on multi-band communication and time division was proposed in [107] to enhance the system performance with respect to energy saving and latency compared to S-MAC protocol. For an ambient assistant living systems, the authors of [108] studied the effects of a distributed time slot scheduling, the channel assignment algorithm and the route establishment on the packet delivery ratio and latency, taking into the cost of switching channel into account. The performance of IEEE 802.11e enhanced distributed channel access (EDCA) was investigated in [109] concerning contention-based differentiated channel access for frames of different priorities in wireless LANs. Moreover, some works focused on the nodes with the capability of sleeping [111], [112]. For example, Liu *et al* in [111] pointed out that alleviated packet accumulation and latency can be achieved for the nodes with a sleep schedule by adaptively adjusting the traffic load measured online. Furthermore, considering that fixed wakeup schedules may not meet the delay constraints when multiple sensors compete for the event delivery at the same time, the authors of [112] proposed a MAC protocol with much lower energy consumption for both single-source and multi-source events. Additionally, since the frame structure plays an important role in determining the system performance of IoT in terms of the delay and reality, investigating the design of frame under the MAC layer is needed [113]. A practical example is that the authors of [114] designed a protocol discarding DATA frames prior to the transmission if the permissible latency cannot be met even if the DATA frame is transmitted. Besides the high efficiency congestion control mechanism design, in order to deal with massive connectivity for HRLL communication in IoT, the low latency grant-free scheduling as introduced below, is a potential solution.

C. Grant-Free Access

With the rapid increase of IoT devices and service, there is increasing pressure for new medium access control (MAC) protocols to support massive and heterogenous IoT devices and services with HRLL requirements for the future wireless networks. Traditionally, reserved time/frequency slots are used to send the schedule request (SR) in order to request resources. Thus, a number of dedicated slots will be occupied by the MAC signaling. Such a MAC protocol is termed as grant-based access (GBA) protocols. However, long round-trip time has to be used for the handshaking in GBA protocols before the granted slots are arranged for the coming payloads transmission. To improve the efficiency and simplify access protocols, grant-free access (GFA) is proposed to avoid the handshaking and waiting delay during the resources scheduling phase. The basic idea of GFA is to permit the UEs randomly choosing the certain slots to send their payloads (information). Thus the time-consuming resources scheduling and allocation overhead/delay is skipped. Clearly, GFA can largely reduce the latency for skipping the scheduling steps and can also reduce the overhead, which is non-trivial for finite block length transmission. Due to collisions caused by possible transmissions of more than one UE over the same slots, the unwanted interference would deteriorate the reliability, or even worse, fail to identify the UEs. By sophisticated design GFA protocols, the receiver could accomplish identification and decoding simultaneously. It has been shown that GFA can achieve higher throughput and much lower latency compared to the original GBA scheme [115].

Several access control schemes are proposed to balance the performance degradation and the system efficiency trade-offs [115]–[119]. The performance comparison between GBA and GFA is studied in [120]–[124]. The results in [125] reveal that GFA can achieve higher throughput under proper retransmission limits. The results of HRLL communications in an outdoor macro scenario are presented in [121], [122], which show that GFA outperforms GBA in terms of latency at the target reliability and achieves the maximum payload with the smallest latency performance degradation [121]. In addition, the asynchronous GFA scheme for short packet communications is investigated in [123], and the reliability, delay, battery lifetime, energy efficiency and spectral efficiency are evaluated as its performance matrices.

Non-orthogonal multiple access (NOMA) has been identified as a promising technology in future IoT systems by sharing the same radio resources (e.g., power domain, code domain). Thus NOMA can enhance the massive connectivity of IoT applications [126]–[130]. Due to the limitations of the conventional random access for massive cellular IoT systems (e.g., preamble collisions and different QoS requirements etc), random NOMA is proposed in [126] and extended to practical considerations. In order to reduce the latency of NOMA, GFA is a nature choice. The NOMA-GFA performance is analyzed and optimized in [127]. Specifically, the analysis of massive grant-free code-domain NOMA is discussed in [128], [129]. Furthermore, pattern division multiple access (PDMA) is applied in the grant-free scheme in [130], in which allocating resource and lowering latency are investigated in details.

To demonstrate the benefit of employing GFA in HRLL IoT networks, we consider the outage probability as a performance metric, which is studied in [120]. To meet the HRLL

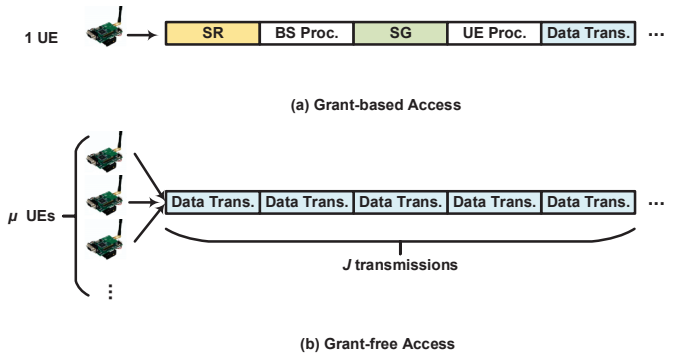

Fig. 14. The models of grant-based and grant-free access: (a) Grant-based access, 1 UE transmission over dedicated channel; (b) Grant-free access, μ UEs share one channel and a target UE gets access by contention, but the collision may be introduced by other ν active UEs.

requirements, we adopt the finite blocklength coding theory with a particularly short packet length, namely, 100 bits per packet to recalculate the outage probability. This is equivalent to evaluate the reliability within a certain latency constraint for GFA. Assume that there is only one channel allowing the uplink access, and the transmission circle is identical for both GBA and GFA, as shown in Fig.14. That is, GBA and GFA consume the same number of transmission time intervals (TTIs) to complete a packet transmission. Clearly for GBA, the UE first sends a scheduling request to the BS. The BS replies to the scheduling request and assigns the dedicated resources to indicate where the payload to be transmitted, and then sends back a scheduling grant (SG) indication signaling to the UE. The time consuming for such a GBA scheme thus is the summation of time for scheduling request phase T_{SR}, BS processing phase T_{BS_P}, scheduling grant phase T_{SG}, processing phase for UE T_{UE_P} and packet transmission phase T_{packet}. Without loss of generality, we assume $T_{SR} = T_{BS\_P} = T_{SG} = T_{UE\_P} = T_{packet} = T_{TTI}$. Thus at least 5 TTIs are needed to finish an interference free packet transmission in one circle for GBA. The error probability of a signaling message, whose errors occur in scheduling request and feedback message, is unified to be the same and denoted as ϵ_s. To compare with GBA, we consider the simplest *blind procedure* GFA, in which the UE transmits its payload ν times in J consecutive TTIs over shared resources without any feedback, in which the transmission could possibly be interfered by other active UEs. Due to interference, errors are unavoidable during the transmission for GFA. Thus there are two types of potential errors. One is the missed UE identification with probability ϵ_I. The other is the data decoding error probability $\epsilon_d(\nu)$ caused by ν other active interfering UEs. An UE is assumed to be equipped with one transmitting antenna and a BS has M receiving antennas, for which the maximum ratio combining (MRC) is assumed. An outage event occurs when the mutual information supported by the instantaneous receive SNR is less than the target data rate, which can be used to evaluate the performance for both GBA and GFA under Rayleigh fading channels with the same latency constraint [120].

- Outage probability for GBA [120]

  The decoding of GBA can be viewed as a collision-free transmission and the probability for correct decoding is

$$P_d^0 = 1 - \epsilon(0). \tag{18}$$

  Considering the possible signaling message error at scheduling request and scheduling grant feedback phases, the outage probability for GBA is

$$P_{outage}^{GBA} = 1 - (1 - \epsilon_s)^2 P_d^0. \tag{19}$$

- Outage probability for GFA [120]

  For the totally J repetition transmission, the outage probability for GFA can be written as

$$P_{outage}^{GFA} = 1 - \sum_{J'=1}^{J} \left(\sum_{j=1}^{J'} \epsilon_I^{j-1} (1-\epsilon_I)^{J'-j+1} (1-P_d)^{J'-j} P_d \right), \tag{20}$$

  where P_d is defined as the probability of correctly decoding a packet

$$P_d = \sum_{\nu=0}^{\mu-1} P_c(\nu)[1 - \epsilon(\nu)], \tag{21}$$

  where $P_c(\nu)$ is given by

$$P_c(\nu) = \binom{\mu - 1}{\nu} P_a^{\nu} (1 - P_a)^{\mu-1-\nu}. \tag{22}$$

Assuming information transmission by each UE follows Poisson distribution, we have $P_a = J(1 - e^{-\lambda})$. For decoding error probability $\epsilon(\mu)$, we will use the formulas for finite blocklength coding theory shown in Section III.A, namely, the normal approximation (10) for the maximal rate and error probability. Considering the interference from ν other UEs, the estimation of error probability $\epsilon(\mu)$ is obtained by averaging over the distribution of the SINR $f_{\rho,\nu}$

$$\epsilon(\nu) = \int_0^{\infty} Q\left(\frac{C(x) - R^*(n,\epsilon)}{\sqrt{\frac{V(x)}{n}}} \right) f_{\rho,\nu}(x) \mathbf{d}x, \tag{23}$$

where $f_{\rho,\nu}(x)$ is the SINR distribution and the BS uses an MRC scheme over flat Rayleigh channel by assuming all the interferers have the same average SNR ρ [131], and

$$\begin{cases} f_{\rho,0}(x) = \dfrac{x^{M-1} e^{-x/\rho}}{\Gamma(M)\rho^M} & \nu = 0, \\ f_{\rho,\nu}(x) = \dfrac{x^{M-1} e^{-x/\rho}}{\Gamma(M)\rho^{M+\nu}} \displaystyle\sum_{m=0}^{M} \binom{M}{m} \frac{\Gamma(m+\nu)\rho^{m+\nu}}{\Gamma(\nu)(x+1)^{m+\nu}} & \nu > 0, \end{cases} \tag{24}$$

where $\Gamma(\cdot)$ is Gamma function. By taking the transmission requirements for FA listed in Table.I, we simulate two example systems to study the performance of grant-free access in terms of outage probability. It is assumed that 20 UEs share a 10MHz bandwidth with packet length 100 bits for both cases. The diversity is set to 1×4 SIMO, i.e., 1 antenna for UE and 4 antennas for the BS. The ϵ_s and ϵ_I are set to be $\min(10^{-5}, 1 - P_d)$.

- LTE-A system

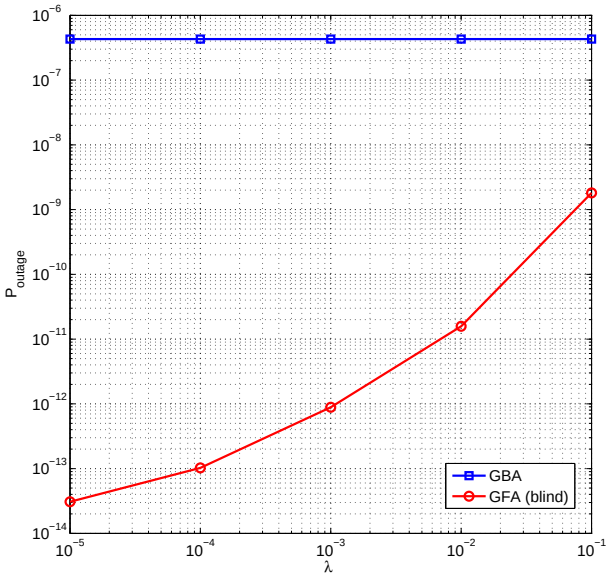

Fig. 15. Outage probability comparison for a mini-slot structure LTE setting, GBA vs GFA.

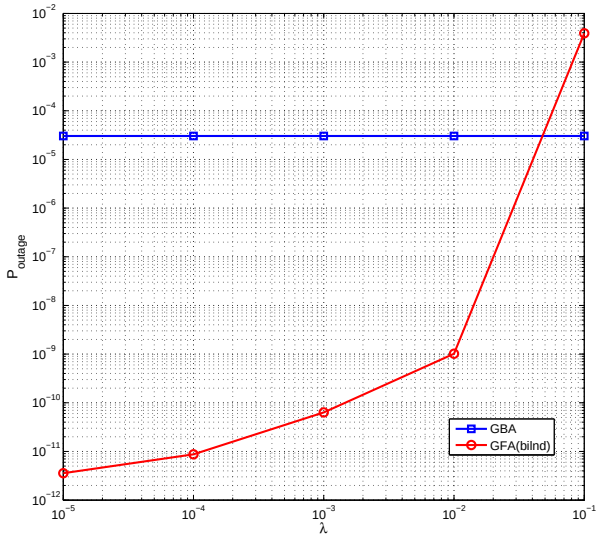

Fig. 16. Outage probability comparison for possible 5G setting for factory automation [134], GBA vs GFA.

For current LTE-A systems [132], the mini-slot structure for possible HRLL usage is defined in [133]. The minimum TTI duration is 2 OFDM symbols and subcarrier space is 15kHz. The outage probabilities that GBA and GFA (blind access) could achieve under different packet arrive ratios are shown in Fig.15. It can be seen clearly that the reliability of GFA would be much better than GBA at the same latency level.

- Possible 5G system
  In [134], a possible system design for 5G FA scenarios is proposed. The objective of the design is to achieve the target PER=10^{-9} within one way delay budget 1 ms. The subcarrier space is 75kHz and an OFDM symbol duration is 14.3μs. The outage probabilities of GBA and GFA are shown in Fig.16. Compared with Fig.15, it can be seen clearly that GFA outperforms GBA at low-to-medium packets arrival rates, but the performance loss can be observed at very high packets arrive rates. This is due to that the scheme in [134] possesses the larger bandwidth over each sub-channel. Thus, in recent discussed 5G standards, the subcarrier space is suggested not to exceed 30kHz [135] for achieving high reliability and low latency.

Although GFA can improve system efficiency theoretically, there are many practical considerations for the technique to be used for IoT, which deserve more attentions [136]–[140]. The GFA cannot be implemented without proper channel estimation and synchronization. Thus it is compulsory to accomplish synchronization, channel estimation, users identification and multi-user detection in a single shot for HRLL transmission [136]. Due to the feature of the sporadic traffic, compressed sensing technique is studied in [137] to formulate user detection and channel estimation problems jointly, and then an efficient approximate message-passing method is employed to improve detection threshold significantly. With massive devices connected in the massive MIMO regime, a two-phase GFA scheme is studied in [138]–[140], in which a BS simultaneously detects the active users and estimates channel by using non-orthogonal pilot sequences in the first phase, while transmits the payload in the second phase. Based on the same system model, achievable transmission rate obtained in [140] can be utilized as the guideline to design the optimum detection method and the corresponding pilot length.

V. Network Layer Analysis and Design for HRLL IoT

In the network layer, latency is a random variable due to various reasons e.g., dynamic traffic or variations in channels. For IoT networks, traffic can be bursty in e.g., accidence or busy hours or lots of working nodes. Thus, it is also quite valuable to consider how network layer impacts the latency of IoT networks. To theoretically analyze the latency in the network layer, network calculus can be used. The main factors impacting latency include traffic, service capability and memory size etc. Among them, traffic and service capability can be optimized by network planning, which includes e.g., network structure and traffic allocation. From the network structure aspect, we can optimize latency by traffic dispersion or network densification etc. Briefly, the traffic dispersion will split the data flow into multiple sub-streams, each of which will be sent through an independent path. Thus, the traffic of each path is reduced. For the network densification, we can densify networks using more nodes, which lead to higher capacity of each path (especially for wireless IoT networks). Then, each node may have higher service capability. Traffic allocation is feasible for more complex networks, in which there are multiple hops from the source to the sink and multiple channels within each hop. More details on network planning are as follows.

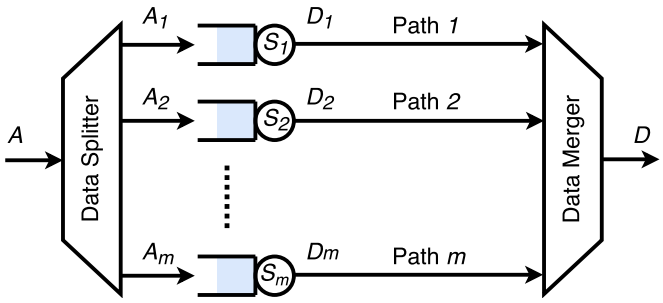

Fig. 17. traffic dispersion.

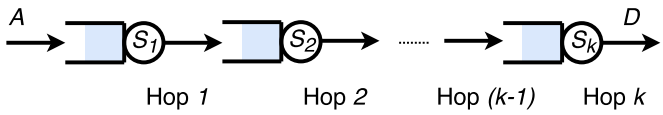

Fig. 18. network densification.

A. Network Structure Optimization

To illustrate the main idea of optimizing the network structure for low latency, let us consider a fluid-flow, discrete-time queuing system with a buffer of infinite size, within time interval $[s,t)$, $0 \le s \le t$, the non-decreasing bivariate processes $A(s,t)$, $D(s,t)$ and $S(s,t)$ are defined as the cumulative arrival traffic to, departure traffic from, and service offered by the server, respectively. We assume $A(s,t)$, $D(s,t)$ and $S(s,t)$ are stationary non-negative random processes, and their values are zeros whenever $s \ge t$. To simplify illustration, we assume a constant arrival rate λ for incoming data traffic. Thus, we have $A(s,t) = \lambda \cdot (t-s)$ for any $0 \le s \le t$. Two schemes of network structuring, i.e., traffic dispersion and network densification can be illustrated as follows.

- For traffic dispersion (as shown in Fig. 17), the arrival traffic is firstly partitioned into multiple sub-streams by the data splitter. More rigorously, given a set of deterministic splitting coefficients $(z_1, z_2, \cdots, z_n)$, where $\sum_{i=1}^{n} z_i = 1$ and $z_i \in (0,1)$ for any $1 \le i \le n$, the i^{th} sub-stream $A_i(s,t)$ is obtained as $A_i(s,t) = z_i \cdot A(s,t)$. Then, each sub-stream is independently served and delivered towards the receiver through the pre-defined path. Finally, the receiver combines all sub-streams through the data merger from different paths, and thereby forming the output traffic. To simplify illustration, we assume independent paths (assuming no interference in the physical layer). Thus, the principle of traffic dispersion is to decompose the original heavy arrival traffic into multiple lighter ones, thereby to avoid a large queue in the buffer, namely, traffic reducing.
- For network densification (as shown in Fig. 18), multiple relay nodes[1] as servers are deployed along the source-destination transmission path. For the concatenation of relying nodes, the output traffic from one relay can be treated as the input traffic for the subsequent connected relay. The application of multi-hop relaying follows certain scenarios of dense wireless IoT networks. Similar to traffic dispersion, it is feasible to assume independent channel conditions on multiple hops. In contrast to traffic dispersion, the mechanism of network densification is to deploy a large number of relay nodes between the given source and destination, which can potentially increase the capacity of each hop via shortening the separation distance between adjacent nodes, thereby increase the end-to-end service capability. The effects of network densification is pronounced for IoT networks since the transmission power of IoT nodes normally is limited due to e.g., limited battery capacity etc. Thus, shortening transmission distance by network densification can significantly increase the channel capacity and thus service capability. Note that the network densification is applied mostly for wireless networks, e.g., major of IoT networks. For wired IoT networks, the analysis and performance of network densification may be quite different.
- Moreover, combining the benefits of traffic dispersion and network densification, hybrid schemes are proposed in [141]. In hybrid schemes, the original arrival traffic is firstly divided into multiple sub-streams by data splitter. Subsequently, these sub-streams are allocated with independent paths for data transmission, and each path consists of multiple relay nodes. It is evident that, this combination takes advantages of traffic dispersion and network densification, i.e., offloading the arrival traffic and enhancing the service capability.

In Fig. 19, we compare the performance of different network planning approaches [141]. As we can see from the figure, for different arrival rates, different schemes will be optimal. For instance, in the low rate regions, densification may be better and in the high rate region, dispersion will be optimal. In the middle rate region, the hybrid rate region may be the best choice. Thus, we should optimize network plan according to expected traffic arrival rates, from the aspect of latency.

B. Traffic Allocation

In addition to optimizing the network structure, we can also improve latency performance by optimizing traffic allocation among channels in the networks, namely, traffic allocation. Consider a multi-hop network with multiple buffer-aided relay nodes and each hop having multiple channels. In the source or the relay nodes, the first-in-first-out (FIFO) rule applies to traffic in the queue of each buffer. Then, traffic allocation will assign the traffic to the individual channels in the source and relay nodes. That is, the traffic at each node (non-sink) is decomposed into a few fractions, which are then pushed into the channels connecting to the next hop, one fraction per channel. The decomposing can continue along the path from the source to the sink. In the relay nodes, the receiving and the transmission can be simultaneous. That is, the relay node work in a full-duplex mode. Traffic allocation has significant impacts to the end-to-end latency for the multi-hop networks with multi-channels in each hop since the traffic congestion at the relay nodes due to non-optimized allocation may lead to long queues and thus large end-to-end latency [142]. Here the end-to-end latency is the time to deliver one fixed-length file

[1]We assume full-duplex relay nodes with negligible self-interference for simplifying network layer analysis.

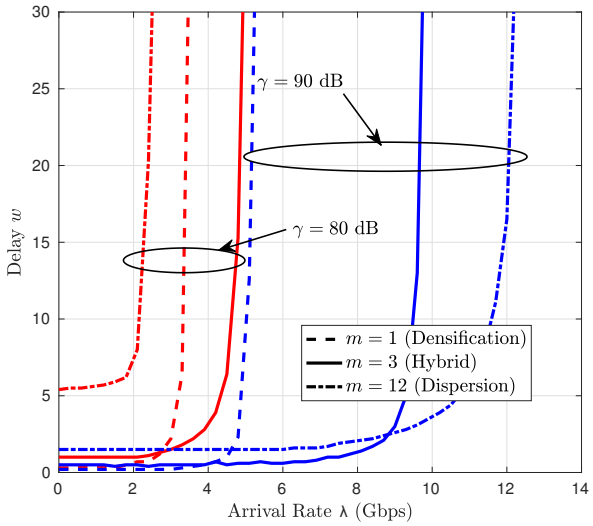

Fig. 19. Probabilistic delay w vs. arrival rate λ for traffic dispersion, network densification and the hybrid scheme, respectively, with respect to violation probability $\epsilon_w = 10^{-3}$, where $n = 12$ and m is the number of independent paths.

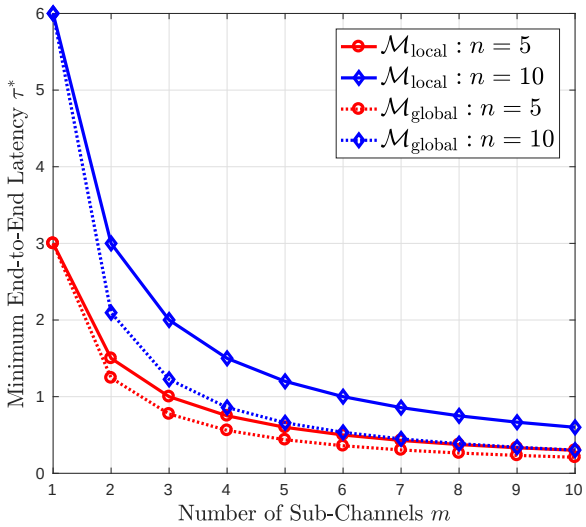

Fig. 20. Minimum end-to-end latency τ^* vs. number of channels m, where the number of relay nodes is $n = 5$ or 10, the capacity of each channel is $C = 1$, and the size of transmitted file is 1 [142].

of size 1. Clearly, the allocation of traffic should be based on the service capability of individual channels, namely, channel capability. In the illustration, we use a tandem networks with n buffer-aided relay nodes and multiple channels in each hop. The source and the destination are denoted as node $n+1$ and node 0, respectively. The hop between the node $h+1$ and h is denoted by hop h, where $0 \le h \le n$. Depending the available channel information, there are roughly two types of traffic allocation schemes [142]: global allocation and local allocation, which are described as follows.

- Local allocation, $\mathcal{M}_{\text{local}}$: Node ζ for all $\zeta \in [n+1]$ only has the capacity information of the channels in hop $\zeta - 1$. The traffic allocation performed at node ζ only optimizes the transmission over channels in hop $\zeta - 1$. This scheme ensures that the latency in the next hop is minimized, but is oblivious to the traffic allocations in other hops.
- Global allocation $\mathcal{M}_{\text{global}}$: Node h for all $h \in [n+1]$ has the entire capacity information of all channels from hop 0 to hop $\zeta - 1$. The traffic allocation performed at node ζ not only relies on channels in hop $\zeta - 1$, but also relies on channels in the remaining hops, i.e., from hop 0 to hop $\zeta - 1$. This scheme minimizes the latency through ζ hops.

To minimize latency, we can formulate the problem to optimize traffic allocation with the objective of end-to-end latency. Actually, in [142], we show that for some special networks, e.g., tandem networks with multiple channels in each hop, the minimum latency for both allocation policies can be derived in a recursive way. In Figure 20, performance comparison between local allocation and global allocation is given. We can see that with more channel information, global allocation has better performance. Meanwhile, if the number of channels per hop increases (higher service capability), the latency will drop quickly.

C. Network Coding

When the latency requirement is high, it may not be preferable to use retransmission schemes, especially in wireless networks, where there may be multiple receivers (transmitters) and transmission errors may occur often. Moreover, in many application scenarios, high data rates are also needed, in addition to high reliability and low latency. For instance, in multimedia dissemination of 5G mobile networks [143] or multi-node industrial control networks, multiple Giga-bps rates may be needed [6]. Yet, in short block coding, meta-data (control information) is non-negligible, which degrades the information-transmission rates. Based on above facts, it is rather interesting to investigate communication and coding schemes capable of simultaneously improving the reliability, latency and rate performance in network layer. Network coding [144], [145], originally proposed to increase the throughput of multicast networks, has shown the benefits of improving latency, throughput and reliability performances in various scenarios. Thus, it is very valuable to consider network coding for HRLL communications. The benefits of applying network coding to HRLL wireless communications include larger rates, reduced latency and higher reliability as detailed below. Moreover, these benefits may be achieved simultaneously with appropriate network coding schemes.

- Larger rate. One major benefit of network coding is to increase information transmission rates. Though the original work of network coding is for multicast networks [144], [145], network coding has shown the advantage of improving rates in various networks, e.g., two-way relay channels [146], [147], multi-hop unicast [148] and multiple unicast networks [149] .
- Reduced latency. Briefly, network coding can directly reduce latency in following ways. Firstly, with higher rates, the transmission latency can be reduced for a given amount of information messages. Secondly, combined

with caching (or storage systems), network coding can substantially reduce transmission latency in some scenarios, e.g., intermediate node caching or terminal caching [150].

- Higher reliability. Network coding has also shown the benefits in increasing reliability in various ways. Firstly, by exploiting path diversity (namely, redundancy from different paths of a network), network coding can be used as network error correction codes [151], which can effectively combat the transmission errors within networks. Secondly, random linear network codes (RLNC) can provide error control capability while used in a rateless [148] way. Compared to fountain codes, RLNC may achieve higher rates and can work efficiently in short blocklength.

Despite these benefits, applying network coding for HRLL communications has mostly not been explored yet. Though there are already many research activities on studying the latency (delay) of network coded systems, the results are still far from sufficiency for HRLL communications. Firstly, the performance for existed results can hardly meet the high requirement of HRLL communications. In [152], the benefits of RLNC are studied in terms of delay and reliability. However, the results in [152] mainly consider the asymptotically results, which assume the coding field size goes to infinity. In [153], the throughput-delay tradeoff is studied for one-hop broadcasting channels with RLNC. The scaling law of throughput relative to the number of users are analyzed. In [154], we study the cross layer optimization problem for minimizing the delay for networks with rateless RLNC. Above results consider the delay problem of coded networks with a large number of packets or infinite coding field size. With these assumptions, the latency may be rather large. Secondly, most of related literatures only show that network coding can improve the delay and reliability performance. Thirdly, the costs of the meta data (packet header) have not been considered yet for coded networks. The costs are non-trivial for short packets. Especially, for RLNC, coding coefficients should be transmitted along with information messages, on top of other meta data. More recently, in [155], network coding is proposed for HRLL communications in industrial control. The benefits of network coding for HRLL communications is clearly shown in a star network [155], which actually exploits the principles of two-way relay channels and cooperative communications. Meanwhile, by exploiting redundancy from path or time, network coding is proposed to increase reliability in various scenarios. In [151], the concept of network error correction codes is proposed and Singleton bound of these codes is shown to be related to the min-cut of the networks. In [148], the error correction capability of linear network coding is explored, in which all source and intermediate nodes store and randomly encode incoming packets, and thus the error correction capability of proposed codes is exploited from both time and path redundancy. After [148], [151], many results discussed the error control capability of network coding. References [152], [153] also discuss the reliability performance of network codes, in addition to delay.

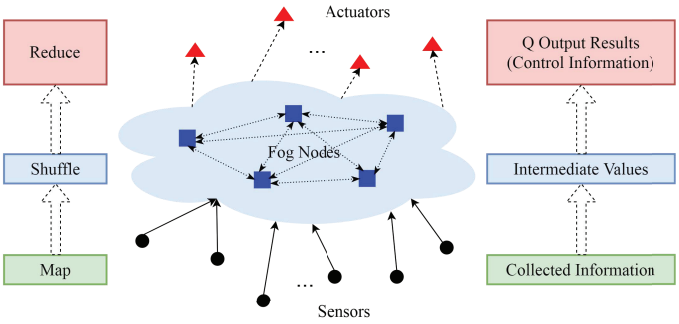

Fig. 21. System model of a BATS-coded fog network with multiple sensors, multiple actuators

The most recent attempt to use network coding in HRLL is in [156]. Herein we considered the fog networks for time-critical applications. Distributed fog computing networks can be used in smart factory, smart city, smart healthcare and smart vehicle for HRLL purpose. Fog network enables the data processing to be distributed and close to the devices, which achieve the high throughput, high reliability and low latency [157] in time-critical control loop: the sensors located around the filed devices collect and transfer the sensed information to fog networks in the uplink for computing and analyzing; in the downlink, the control commands are generated and sent to actuators from fog nodes. However, the overall response latency could be deteriorated by high communication load, due to a huge amount of data exchanging among fog nodes. Thus, an enhanced rateless sparse RLNC scheme, termed as batched sparse (BATS) codes [158], [159], was proposed for transmitting a collection of data through fog networks, which leads to less coding overhead and lower latency. A system model for BATS-coded fog networks is shown in Fig.21. S sensors BATS-encode their sensed data and send the coded packets to the F fog nodes for processing, then the calculated intermediate values are BATS-encoded again and exchanged among fog nodes in the Data Shuffling stage, and finally Q control commands are generated and sent to the A actuators after BATS-decoding in the Reduce stage. Due to transmission loss or uncompleted computation, the channels can be regarded as erasure ones. The procedure of employing BATS codes is following:

1) Mapping Stage: Suppose there are total M nodes, and K collected data packets are BATS encoded by each sensor to form n batches, and g fog nodes are assigned to compute one common batch.
2) Data Shuffling Stage: One subset of fog nodes receives N packets from the ith batch and Q intermediate values are generated at each fog node. Then, each fog node performs BATS encoding on the $\frac{NQ}{g}$ intermediate values to form $\frac{MQ}{g}$ new encoded packets which are exchanged among the fog nodes in different subsets.
3) Reduce Stage: After collecting enough K intermediate values, the fog nodes BATS-decode the data, calculate the corresponding control command and send to actuators.

Several works have considered to use RLNC in fog networks to lower the communication load and latency [160], [161]. We use these results as our comparison baseline. In [156], the overall response time for RLNC-coded $L_{\text{overall, coded}}$ and BATS-coded $L_{\text{overall,bats}}$ are derived and analyzed (refers to (5) and (6) of [156]). If the computation loads of the RLNC-coded and BATS-coded schemes are set to be the same values, i.e., r, the different of the overall response time will be mainly determined by the communication loads. Therefore, compared with the coded scheme, the ratio of time reduction over RLNC-coded scheme by the BATS-coded scheme can be written as

$$R_{\text{coded,bats}} = \left(1 - \frac{L_{\text{overall,bats}}(r-1,g)}{L_{\text{overall, coded}}(r)}\right) \times 100\%. \quad (25)$$

Fig. 22 shows an example of the ratio of time reduction by using the BATS-coded scheme when r varies from 1 to 10, $F = 10$ fog nodes, $\frac{M}{K} = 0.1$, $g = 1$, and the erasure probability e_F changes from 0.1 to 0.4. As the channel condition becomes worse and/or the computation load becomes lower, e.g., e_F increases from 0.1 to 0.4, r decreases from 6 to 1, more time for exchanging the intermediate results can be reduced by using the BATS-coded scheme. Although RLNC coded scheme has already reduce the response time a lot [160], [161], the shorter overall response time can be achieved by the BATS-coded scheme over the RLNC-coded one.

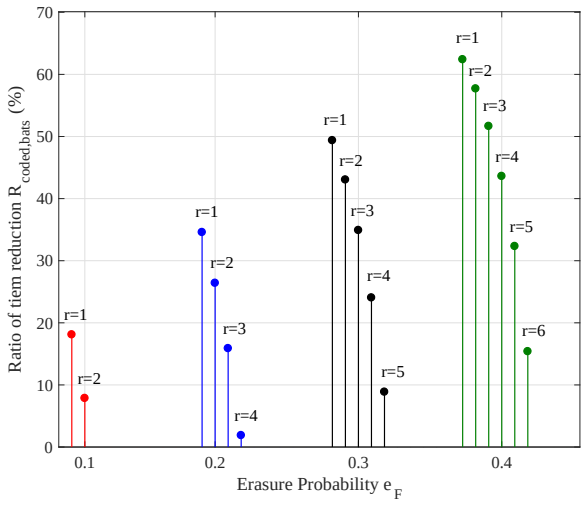

Fig. 22. An example of the ratio of time reduction by using the BATS-based scheme with r.

VI. Concluding Remarks

With the advances in information theory and wireless communication techniques and emerging applications, HRLL IoT research has attracted lots of attention recently. We have surveyed typical application scenarios, techniques and various possible solutions for HRLL IoT networks. For the different requirements on the code length, there are often tradeoffs between latency and reliability. Thus, it is rather challenging to achieve high-reliability and low latency simultaneously, especially for resource limited IoT networks. To achieve HRLL requirements, we need to optimize various strategies, which impact frame size, preamble, network/channel coding, multiple access, resource scheduling, network optimization etc. Especially for some critical applications, such as power systems automation, power electronics control and factory automation, sophisticated designs should be performed. Future work on this may include the designing practical schemes and systems for various scenarios according to their different requirements. Significant work still remains ahead in HRLL IoT networks, ranging from physical layer (e.g., packet structure optimization, preamble/pilots design, massive MIMO, short and high reliable channel coding, synchronization, channel modeling and estimation, etc) to MAC and network layer (e.g. initial access, mmWave transmission, non-orthogonal multiple access etc). As one of the most stringent requirements to be supported in future radio transmission, HRLL communication will be able to support a number of vertical industries. To that end, we foresee that HRLL communication will play a key role in future IoT networks.

References

[1] "HART field communication protocol specification, revision 7.0," *HART Communication Foundation Std., [Online]. Available: http://www.hartcomm.org/*, 2007.

[2] G. Scheible, D. Dzung, J. Endresen, and J. E. Frey, "Unplugged but connected-design and implementation of a truly wireless real-time sensor/actuator interface," *IEEE Industrial Electronics Magazine*, vol. 1, no. 2, pp. 25–34, July 2007.

[3] "IEC PAS 62601: Industrial networks-wireless communication network and communication profiles–WIA-PA," *International Electrotechnical Commission (IEC) Std.*, 2015.

[4] "IEC PAS 62948: Industrial networkswireless communication network and communication profiles–WIA-FA," *International Electrotechnical Commission (IEC) Std.*, 2015.

[5] "802.11-IEEE standard for information technologylocal and metropolitan area networks. part 15: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std.*, 2012.

[6] M. Luvisotto, Z. Pang, and D. Dzung, "Ultra high performance wireless control for critical applications: Challenges and directions," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1448–1459, March, 2016.

[7] M. Zhan, Z. Pang, D. Dzung, and M. Xiao, "Channel coding for high performance wireless control in critical applications: Survey and analysis," *IEEE Transactions on Industrial Informatics*, To Appear.

[8] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

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

[10] P. Popovski, J. J. Nielsen, C. Stefanovic, E. D. Carvalho, E. Strom, K. F. Trillingsgaard, A. S. Bana, M. K. Dong, R. Kotaba, and J. Park, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[11] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[12] A. Nasrallah, A. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research," *IEEE Communications Surveys & Tutorials*, vol. to be published, pp. 1–59, 2019.

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

[14] T. Richardson and S. Kudekar, "Design of low-density parity check codes for 5G new radio," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 28–34, March 2018.

[15] Z. Pang, M. Luvisotto, and D. Dzung, "Wireless high-performance communications: The challenges and opportunities of a new target," *IEEE Industrial Electronics Magazine*, vol. 11, no. 3, pp. 20–25, September 2017.

[16] M. Luvisotto, Z. Pang, and D. Dzung, "Ultra high performance wireless control for critical applications: Challenges and directions," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1448–1459, June 2017.

[17] J. Q. Li, F. R. Yu, G. Deng, C. Luo, and Q. Yan, "Industrial internet: A survey on the enabling technologies, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1504–1526, July 2017.

[18] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, March 2017.

[19] X. Wang, S. Mao, and M. X. Gong, "An overview of 3GPP cellular vehicle-to-everything standards," *Getmobile Mobile Computing & Communications*, vol. 21, no. 3, pp. 19–25, November 2017.

[20] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2377–2396, October 2015.

[21] A. Adnan and M. Sooriyabandara, "The tactile internet for industries: A review," *The Procedding of IEEE*, vol. to be published, pp. 1–22, February 2019.

[22] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5g-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.

[23] J. Mosyagin, "Using 4G wireless technology in the car," in *International Conference on Transparent Optical Networks*, July 2010, pp. 1–4.

[24] G. Araniti, C. Campolo, M. Condoluci, and A. Iera, "LTE for vehicular networking: a survey," *IEEE Communications Magazine*, vol. 51, no. 5, pp. 148–157, May 2013.

[25] A. N. Kim, F. Hekland, S. Petersen, and P. Doyle, "When HART goes wireless: Understanding and implementing the wirelessHART standard," in *IEEE International Conference on Emerging Technologies & Factory Automation*, September 2008.

[26] A. Willig, K. Matheus, and A. Wolisz, "Wireless Technologies in Industrial Networks," *Proceedings of the IEEE*, vol. 93, no. 6, pp. 1130–1150, June 2005.

[27] D. Orfanus, R. Indergaard, G. Prytz, and T. Wien, "EtherCAT-based platform for distributed control in high-performance industrial applications," in *2013 IEEE 18th Conference on Emerging Technologies Factory Automation (ETFA)*, September 2013, pp. 1–8.

[28] H. Zhu, Z. Pang, B. Xie, and G. Bag, "IETF IoT based wireless communication for latency-sensitive use cases in building automation," in *2016 IEEE 25th International Symposium on Industrial Electronics (ISIE)*, June 2016, pp. 1168–1173.

[29] K. Yu, Z. Pang, M. Gidlund, J. A.kerberg, and M. Bjorkman, "REALFLOW: Reliable real-time flooding-based routing protocol for industrial wireless sensor networks," *Int. J. Distrib. Sensor Netw*, vol. 2014, pp. 936 379–1936 379–17, July 2014.

[30] S. Feliciano, H. Sarmento, and J. de Oliverira, "Field area network in a MV/LV substation: A technical and economical analysis," in *2014 IEEE International Conference on Intelligent Energy and Power Systems (IEPS)*, June 2014, pp. 192–197.

[31] D. Cottet and *et. al*, "Integration technologies for a fully modular and hot-swappable MV multi-level concept converter," in *Proceedings of PCIM Europe 2015; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management*, May 2015, pp. 1–8.

[32] "http://www.3gpp.org."

[33] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, and W. Xu, "Use cases, requirements, and design considerations for 5G V2X," *IEEE Vehicular Technology Magazine*, preprint, 2017.

[34] A. Festag, "Cooperative intelligent transport systems standards in europe," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 166–172, December 2014.

[35] "3GPP TR 22.886 V1.0.0: Study on enhancement of 3GPP support for 5G V2X services (Release 15)," 2016.

[36] "ETSI TC ITS, intelligent transport systems (ITS); vehicular communications; basic set of applications," 2015.

[37] M. Kuzlu, M. Pipattanasomporn, and S. Rahman, "Communication network requirements for major smart grid applications in HAN, NAN and WAN," *Computer Networks*, vol. 67, no. 10, pp. 74–88, July 2014.

[38] D. Burmester, R. Rayudu, W. Seah, and D. Akinyele, "A review of nanogrid topologies and technologies," *Renewable & Sustainable Energy Reviews*, vol. 67, pp. 760–775, January 2017.

[39] A. Riccobono, M. Ferdowsi, J. Hu, H. Wolisz, P. Jahangiri, D. Mller, R. W. D. Doncker, and A. Monti, "Next generation automation architecture for DC smart homes," in *Energy Conference*, April 2016, pp. 1–6.

[40] L. K. Siow, P. L. So, H. B. Gooi, F. L. Luo, C. J. Gajanayake, and Q. N. Vo, "Wi-Fi based server in microgrid energy management system," in *TENCON 2009 - 2009 IEEE Region 10 Conference*, January 2009, pp. 1–5.

[41] V. Tanyingyong, R. Olsson, J. W. Cho, M. Hidell, and P. Sjodin, "IoT-grid: IoT communication for smart dc grids," in *Global Communications Conference*, December 2016.

[42] "ETSI TS: Machine to machine communications (M2M): M2M service requirements," 2010.

[43] Y. Kabalci, "A survey on smart metering and smart grid communication," *ELSEVIER Renewable and Sustainable Energy Reviews*, vol. 57, pp. 302–318, May 2016.

[44] C. Marnay, B. Nordman, and J. Lai, "Future Roles of Milli-, Micro-, and Nano- Grids," *in Proc. of CIGRE International Symposium*, 2011.

[45] M. Sechilariu, B. Wang, and F. Locment, "Building integrated photovoltaic system with energy storage and smart grid communication," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 4, pp. 1607–1618, April 2013.

[46] D. Q. Xu, G. Joos, M. Levesque, and M. Maier, "Integrated V2G, G2V, and renewable energy sources coordination over a converged fiber-wireless broadband access network," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1381–1390, Sep 2013.

[47] J. Taylor, A. Maitra, M. Alexander, D. Brooks, and M. Duvall, "Evaluation of the impact of plug-in electric vehicle loading on distribution system operations," in *Power & Energy Society General Meeting, 2009. PES '09. IEEE*, July 2009, pp. 1–6.

[48] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, July 1948.

[49] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[50] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[51] G. Durisi, T. Koch, J. stman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 618–629, February 2016.

[52] K. F. Trillingsgaard and P. Popovski, "Downlink transmission of short packets: Framing and control information revisited," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2048–2061, May 2017.

[53] B. Lee, S. Park, D. J. Love, H. Ji, and B. Shim, "Packet structure and receiver design for low latency wireless communications with ultra-short packets," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 796–807, February 2017.

[54] G. Liva, G. Durisi, M. Chiani, S. S. Ullah, and S. C. Liew, "Short codes with mismatched channel state information: A case study," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications*, July 2017, pp. 1–5.

[55] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai Shitz, "On information rates for mismatched decoders," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1953–1967, November 1994.

[56] S. J. Johnson, *Iterative Error Correction, Turbo, Low-Density Parity-Check and Repeat-Accumulate Codes*. Cambridge University Press, 2019.

[57] Z. Ma, W. H. Mow, and P. Fan, "On the complexity reduction of turbo decoding for wideband CDMA," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 353–356, March 2005.

[58] Z. Ma, P. Fan, W. H. Mow, and Q. Chen, "A joint early detection-early stopping scheme for short-frame turbo decoding," *AEUE - International Journal of Electronics and Communications*, vol. 65, no. 1, pp. 37–43, January 2011.

[59] B. Y. Chang, D. Divsalar, and L. Dolecek, "Non-binary protograph-based LDPC codes for short block-lengths," in *2012 IEEE Information Theory Workshop*, September 2012, pp. 282–286.

[60] V. Rybalkin, P. Schlafer, and N. Wehn, "A new architecture for high speed, low latency NB-LDPC check node processing for GF(256)," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[61] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[62] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[63] K. Niu and K. Chen, "CRC-aided decoding of polar codes," *IEEE Communications Letters*, vol. 16, no. 10, pp. 1668–1671, October 2012.

[64] K. Niu, K. Chen, J. Lin, and Q. T. Zhang, "Polar codes: Primary concepts and practical decoding algorithms," *IEEE Communications Magazine*, vol. 52, no. 7, pp. 192–203, July 2014.

[65] D. Wu, Y. Li, X. Guo, and Y. Sun, "Ordered statistic decoding for short polar codes," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1064–1067, June 2016.

[66] K. Niu, K. Chen, and J. R. Lin, "Beyond turbo codes: Rate-compatible punctured polar codes," in *2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 3423–3427.

[67] K. Niu, K. Chen, and J. Lin, "Low-complexity sphere decoding of polar codes based on optimum path metric," *IEEE Communications Letters*, vol. 18, no. 2, pp. 332–335, February 2014.

[68] O. Dizdar and E. Arkan, "A high-throughput energy-efficient implementation of successive cancellation decoder for polar codes using combinational logic," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 3, pp. 436–447, March 2016.

[69] S. M. Abbas, Y. Fan, J. Chen, and C. Y. Tsui, "High-throughput and energy-efficient belief propagation polar code decoder," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 3, pp. 1098–1111, March 2017.

[70] Y. Lee, H. Yoo, I. Yoo, and I. C. Park, "High-throughput and low-complexity BCH decoding architecture for solid-state drives," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 5, pp. 1183–1187, May 2014.

[71] Y. H. Yitbarek, K. Yu, J. Åkerberg, M. Gidlund, and M. Björkman, "Implementation and evaluation of error control schemes in industrial wireless sensor networks," in *2014 IEEE International Conference on Industrial Technology (ICIT)*, March 2014, pp. 730–735.

[72] C. Rachinger, J. B. Huber, and R. R. Mller, "Comparison of convolutional and block codes for low structural delay," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4629–4638, December 2015.

[73] D. S. Yoo, W. E. Stark, K. P. Yar, and S. J. Oh, "Coding and modulation for short packet transmission," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 2104–2109, May 2010.

[74] C. Studer, S. Fateh, C. Benkeser, and Q. Huang, "Implementation trade-offs of soft-input soft-output MAP decoders for convolutional codes," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 11, pp. 2774–2783, November 2012.

[75] T. Hehn and J. B. Huber, "LDPC codes and convolutional codes with equal structural delay: a comparison," *IEEE Transactions on Communications*, vol. 57, no. 6, pp. 1683–1692, June 2009.

[76] N. ul Hassan, M. Lentmaier, and G. P. Fettweis, "Comparison of LDPC block and LDPC convolutional codes based on their decoding latency," in *2012 7th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, August 2012, pp. 225–229.

[77] S. V. Maiya, D. J. Costello, T. E. Fuja, and W. Fong, "Coding with a latency constraint: The benefits of sequential decoding," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, October 2010, pp. 201–207.

[78] M. M. Kermani, V. Singh, and R. Azarderakhsh, "Reliable low-latency Viterbi algorithm architectures benchmarked on ASIC and FPGA," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 1, pp. 208–216, January 2017.

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

[80] E. Kalantari, I. Boryaliniz, A. Yongacoglu, and H. Yanikomeroglu, "User association and bandwidth allocation for terrestrial and aerial base stations with backhaul considerations," in *2012 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Montreal, QC, Canada, October 2017, pp. 1–6.

[81] L. Chen, B. Chang, G. Zhao, and Z. Chen, "Delay-sensitive area spectral efficiency optimization for uplink transmission in ultra-reliable and low-latency communications," in *23rd Asia-Pacific Conference on Communications (APCC 2017)*, Perth, WA, Australia, December 2017, pp. 1–6.

[82] M. Li, X. Guan, C. Hua, C. Chen, and L. Lyu, "Predictive pre-allocation for low-latency uplink access in industrial wireless networks," http://cn.arxiv.org/pdf/1801.06451v1.

[83] B. Singh, O. Tirkkonen, Z. Li, M. Uusitalo, B. Singh, O. Tirkkonen, Z. Li, and M. Uusitalo, "Resource allocation strategy for ultra-reliable communication in a factory environment," in *Global Wireless Summit*, Aarhus, Denmark, November 2016, pp. 1–6.

[84] H. Vesa, Z. Li, B. Soret, and V. Nurmela, "Coordinated multi-cell resource allocation for 5g ultra-reliable low latency communications," in *European Conference on Networks and Communications*, Oulu, Finland, June 2017, pp. 1–5.

[85] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, and Z. Zhang, "Enabling ultra-reliable and low-latency communications through unlicensed spectrum," *IEEE Network*, vol. 32, no. 2, pp. 70–77, 2018.

[86] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *IEEE Global Communications Conference (GLOBECOM 2017)*, Singapore, Singapore, December 2017, pp. 1–6.

[87] M. I. Ashraf, C. F. Liu, M. Bennis, and W. Saad, "Towards low-latency and ultra-reliable vehicle-to-vehicle communication," in *European Conference on Networks and Communications (EuCNC 2017)*, Oulu, Finland, June 2017, pp. 1–6.

[88] C. F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1292–1295, June 2018.

[89] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," http://cn.arxiv.org/pdf/1805.01332v1.

[90] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated learning for ultra-reliable low-latency v2v communications," http://cn.arxiv.org/pdf/1805.09253v1.

[91] T. K. Vu, C. F. Liu, M. Bennis, M. Debbah, and M. Latva-Aho, "Path selection and rate allocation in self-backhauled mmwave networks," in *IEEE Wireless Communications and Networking Conference (WCNC 2018)*, Barcelona, Spain, April 2018, pp. 1–6.

[92] T. K. Vu, C. F. Liu, M. Bennis, M. Debbah, M. Latva-Aho, and C. S. Hong, "Radio resource and traffic management for ultra-reliable low latency communications," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Spain, April 2018, pp. 1–6.

[93] T. K. Vu, C. F. Liu, M. Bennis, and M. Debbah, "Ultra-reliable and low latency communication in mmwave-enabled massive mimo networks," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2041–2044, 2017.

[94] C. F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: an extreme value theory approach," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1292–1295, June 2018.

[95] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Virtual resource allocation for heterogeneous services in full duplex-enabled scns with mobile edge computing and caching," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1794–1808, February 2018.

[96] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Grant-free transmissions for ultra-reliable and low latency uplink communications," in *European Conference on Networks and Communications (EUCNC 2018)*, Ljubljiana, Slovenia, June 2018, pp. 1–6.

[97] R. Jurdi, S. R. Khosravirad, and H. Viswanathan, "Variable-rate ultra-reliable and low-latency communication for industrial automation," in *Conference on Information Sciences and Systems*, Princeton, NJ, United states, March 2018, pp. 1–6.

[98] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer transmission design for tactile internet," in *59th IEEE Global Communications Conference (GLOBECOM 2016)*, Washington, DC, United states, December 2016, pp. 1–6.

[99] N. A. Johansson, Y. P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5g communications," in *IEEE International Conference on Communication Workshop*, London, United kingdom, June 2015, pp. 1184–1189.

[100] J. Oestman, G. Durisi, E. G. Stroem, J. Li, H. Sahlin, and G. Liva, "Low-latency ultra-reliable 5g communications: finite-blocklength bounds and coding schemes," in *SCC 2017; International Itg Conference on Systems, Communications and Coding; Proceedings of*, Hamburg, Germany, February 2017, pp. 1–6.

[101] M. Serror, V. Sebastian, K. Wehrle, and J. Gross, "Practical evaluation of cooperative communication for ultra-reliability and low-latency," in *19th IEEE International Symposium on A world of Wireless, Mobile, Mobile and Multimedia Network (WOWMOM 2018)*, Chania, Greece, June 2018, pp. 1–6.

[102] C. Sun, C. She, and C. Yang, "Exploiting multi-user diversity for ultra-reliable and low-latency communications," in *IEEE GLOBECOM Workshops*, Singapore, Singapore, December 2017, pp. 1–6.

[103] Y. L. Gao, Y. Xiao, M. M. Wu, M. Xiao, and J. L. Shao, "Dynamic social-aware peer selection for cooperative relay management with d2d communications," to be published.

[104] H. Jung, S. G. Kim, H. Y. Yeom, S. Kang, and L. Libman, "Adaptive delay-based congestion control for high bandwidth-delay product networks," in *IEEE INFOCOM*, Shanghai, China, April 2011, pp. 1–6.

[105] M. Bahnasy, H. Elbiaze, and B. Boughzala, "HetFlow: A distributed delay-based congestion control for data centers to achieve ultra low queueing delay," in *IEEE Int Conf Commun*, Paris, France, May 2017, pp. 1–6.

[106] S. D'Aronco, L. Toni, S. Mena, X. Zhu, and P. Frossard, "Improved utility-based congestion control for delay-constrained communication," *IEEE/ACM Transactions on Networking*, vol. 25, no. 1, pp. 349–362, February 2015.

[107] H. Zayani, R. B. Ayed, K. Djouani, and K. Barkaoui, "ECO-MAC: an energy-efficient and low-latencyhybrid MAC protocol for wireless sensor networks," in *ACM Workshop on Performance Monitoring & Measurement of Heterogeneous Wireless & Wired Networks*. Chania, Crete Island, Greece: IEEE, October 2007, pp. 1–4.

[108] H. Chen and L. Cui, "DS-MMAC:a delay-sensitive multi-channel MAC protocol for ambient assistant living systems," *China Communications*, vol. 13, no. 5, pp. 38–46, May 2016.

[109] C. L. Huang and W. Liao, "Throughput and delay performance of IEEE 802.11e enhanced distributed channel access (EDCA) under saturation condition," *IEEE Transactions on Wireless Communications*, vol. 6, no. 1, pp. 136–145, January 2007.

[110] J. W. Tantra, C. H. Foh, and A. B. Mnaouer, "Throughput and delay analysis of the IEEE 802.11e edca saturation," in *IEEE International Conference on Communications*, Seoul, South Korea, May 2005, pp. 1–6.

[111] H. Liu, G. Yao, J. Wu, and L. Shi, "An energy-efficient and low-latency MAC protocol for wireless sensor networks," *J. Commun. Netw.*, vol. 12, no. 5, October 2010.

[112] L. Choi, S. H. Lee, and J. Jun, "SPEED-MAC: Speedy and energy efficient data delivery MAC protocol for real-time sensor network applications," in *IEEE Int. Conf. Commun.*, Cape Town, South Africa, July 2010, pp. 1–6.

[113] S. S. Joo, B. S. Kim, J. A. Jun, and C. S. Pyo, "Enhanced MAC for the bounded access delay," in *Int. Conf. Inf. Commun. Technol. Convergence*, Jeju, South Korea, November 2010, pp. 1–6.

[114] S. Konishi and X. Wang, "A MAC protocol taking account of permissible latency for real-time applications," in *IEEE Int. Symp. Person Indoor Mobile Radio Commun.*, Athens, Greece, September 2007, pp. 1–6.

[115] S. Chae, S. Cho, S. Kim, and M. Rim, "Coded random access with multiple coverage classes for massive machine type communication," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, October 2016, pp. 882–886.

[116] S. Hu, H. Guo, C. Jin, Y. Huang, B. Yu, and S. Li, "Frequency-domain oversampling for cognitive CDMA systems: Enabling robust and massive multiple access for internet of things," *IEEE Access*, vol. 4, pp. 4583–4589, August 2016.

[117] H. Zhang, R. Li, J. Wang, Y. Chen, and Z. Zhang, "Reed-muller sequences for 5G grant-free massive access," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, December 2017, pp. 1–7.

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

[119] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive multiple access based on superposition raptor codes for cellular M2M communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 307–319, January 2017.

[120] G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovcs, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23 602–23 611, April 2018.

[121] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System level analysis of uplink grant-free transmission for URLLC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, December 2017, pp. 1–6.

[122] C. Wang, Y. Chen, Y. Wu, and L. Zhang, "Performance evaluation of grant-free transmission for uplink URLLC services," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–6.

[123] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *2017 IEEE Global Communications Conference (GLOBECOM)*, December 2017, pp. 1–7.

[124] W. Yu, "On the fundamental limits of massive connectivity," in *2017 Information Theory and Applications Workshop (ITA)*, February 2017, pp. 1–6.

[125] C. Pyo, K. Takizawa, M. Moriyama, M. Oodo, H. Tezuka, K. Ishizu, and F. Kojima, "A throughput study of grant-free multiple access for massive wireless communications," in *2017 20th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, December 2017, pp. 529–534.

[126] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, September 2017.

[127] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2238–2252, October 2017.

[128] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "On the performance of massive grant-free NOMA," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, October 2017, pp. 1–6.

[129] ——, "Grant-free massive NOMA: Outage probability and throughput," *arXiv*, no. 1707.07401, July 2017.

[130] W. Tang, S. Kang, B. Ren, and X. Yue, "Uplink grant-free pattern division multiple access for 5G radio access," *China Communications*, vol. 15, no. 4, pp. 153–163, April 2018.

[131] V. A. Aalo and J. Zhang, "Performance analysis of maximal ratio combining in the presence of multiple equal-power cochannel interferers in a nakagami fading channel," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 2, pp. 497–503, March 2001.

[132] D. H. Holma and D. A. Toskala, *LTE for UMTS - OFDMA and SC-FDMA Based Radio Access*. Wiley Publishing, 2009.

[133] "3GPP TR 38.802 V14.2.0: Study on new radio access technology-physical layer aspects," 2017.

[134] O. N. C. Yilmaz, Y. . 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 *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 1190–1195.

[135] "3GPP TR 38.211 V15.2.0: Physical channels and modulation," 2018.

[136] A. T. Abebe and C. G. Kang, "Comprehensive grant-free random access for massive amp; low latency communication," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[137] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1890–1904, April 2018.

[138] L. Liu and W. Yu, "Massive device connectivity with massive MIMO," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1072–1076.

[139] ——, "Massive connectivity with massive MIMOPart I: Device activity detection and channel estimation," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2933–2946, June 2018.

[140] ——, "Massive connectivity with massive MIMOPart II: Achievable rate characterization," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2947–2959, June 2018.

[141] G. Yang, M. Xiao, and H. V. Poor, "Low-latency millimeter-wave communications: Traffic dispersion or network densification?" *IEEE Transactions on Communications*, vol. 66, no. 8, pp. 3526–3539, August 2018.

[142] G. Yang, M. Haenggi, and M. Xiao, "Traffic allocation for low-latency multi-hop networks with buffers," *IEEE Transactions on Communications*, vol. 66, no. 9, pp. 3999 – 4013, September 2018.

[143] J. G. Andrews, S. Buzzi, C. Wan, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June, 2014.

[144] R. Ahlswede, N. Cai, S. Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, April, 2000.

[145] S. Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 371–381, February, 2003.

[146] S. Zhang, S. C. Liew, and P. P. Lam, "Hot topic: Physical-layer network coding," in *International Conference on Mobile Computing & Networking*, January 2006.
[147] P. Popovski and H. Yomo, "Physical network coding in two-way wireless relay channels," *Proc IEEE Icc Glasgow Scotland*, pp. 707–712, June 2007.
[148] D. S. Lun, M. Mdard, R. Koetter, and M. Effros, "Further results on coding for reliable communication over packet networks." *Physical Communication*, vol. 1, no. 1, pp. 3–20, January, 2008.
[149] B.-Y. Ziv, J. Yitzhak, T. S., and K. Tomer, "Index coding with side information." *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1479–1494, March 2011.
[150] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 1077–1081, May, 2014.
[151] R. W. Yeung and N. Cai, "Network error correction, I: Basic concepts and upper bounds," *Communications in Information & Systems*, vol. 6, no. 1, p. 2006, January, 2006.
[152] A. Eryilmaz, A. Ozdaglar, M. Medard, and E. Ahmed, "On the delay and throughput gains of coding in unreliable networks," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5511–5524, December, 2008.
[153] B. T. Swapna, A. Eryilmaz, and N. B. Shroff, "Throughput-delay analysis of random linear network coding for wireless broadcasting," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6328–6341, October, 2013.
[154] X. Ming, "Cross-layer design of rateless random network codes for delay optimization," *IEEE Transactions on Communications*, vol. 59, no. 12, pp. 3311–3322, December, 2011.
[155] V. N. Swamy, P. Rigge, G. Ranade, A. Sahai, and B. Nikolic, "Network coding for high-reliability low-latency wireless control," in *Wireless Communications & Networking Conference Workshops*, April 2016.
[156] Y. Jing, X. Ming, and Z. Pang, "Distributed fog computing based on batched sparse codes for industrial control," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4683–4691, October, 2018.
[157] A. C. G. Juan, E. L. R. Daniel, and F. H. P. Fitzek, "On network coded distributed storage: How to repair in a fog of unreliable peers," in *International Symposium on Wireless Communication Systems*, September 2016.
[158] S. Yang and R. W. Yeung, "Batched sparse codes," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5322–5346, September, 2014.
[159] T. C. Ng and S. Yang, "Finite-length analysis of bats codes," in *International Symposium on Network Coding*, January 2013.
[160] S. Li, Y. Qian, M. A. Maddah-Ali, and A. S. Avestimehr, "Coded distributed computing: Fundamental limits and practical challenges," in *Conference on Signals, Systems & Computers*, March 2017.
[161] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental tradeoff between computation and communication in distributed computing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 109–128, January 2018.