Enabling Grant-Free URLLC: An Overview of Principle and Enhancements by Massive MIMO

Jie Ding, Mahyar Nemati, Shiva Raj Pokhrel, Ok-Sun Park, Jinho Choi, *Senior Member, IEEE*, and Fumiyuki Adachi, *Life Fellow, IEEE*

Abstract—**Enabling ultra-reliable low-latency communication (URLLC) with stringent requirements for transmitting data packets (e.g., 99.999% reliability and 1 millisecond latency) presents considerable uplink transmission challenges. For each packet transmission over dynamically allocated network radio resources, the conventional random access protocols are based on a request-grant scheme. This induces excessive latency and necessitates reliable control signalling, resulting in overhead. To address these problems, grant-free (GF) solutions are proposed in the fifth-generation (5G) new radio (NR). In this paper, an overview and vision of the state-of-the-art in enabling GF URLLC are presented. In particular, we first provide a comprehensive review of NR specifications and techniques for URLLC, discuss underlying principles, and highlight impeding issues of enabling GF URLLC. Furthermore, we briefly explain two key phenomena of massive multiple-input multiple-output (mMIMO) (i.e., channel hardening and favorable propagation) and build several deep insights into how celebrated mMIMO features can be exploited to address the issues and enhance the performance of GF URLLC. Moving further ahead, we examine the potential of cell-free (CF) mMIMO and analyze its distinctive features and benefits over mMIMO to resolve GF URLLC issues. Finally, we identify future research directions and challenges in enabling GF URLLC with CF mMIMO.**

Index Terms—**URLLC, grant-free random access, retransmission, massive MIMO, cell-free.**

I. Introduction

The excitement about fifth-generation (5G) is ushering in the possibility of dramatic network improvements and motivating service providers to plan for the future [1]. Today's Internet of Things (IoT) and networks support a broad range of services by providing ubiquitous all-purpose connectives [2]. 5G networks are expected to support wireless connections originated by various IoT devices in addition to traditional broadband services. To this end, three new service categories were defined by the 3rd generation partnership project (3GPP) for 5G [3], namely enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low-latency communication (URLLC). Among them, enabling URLLC service is the most challenging task in 5G and future wireless networks as two ambitious requirements of high reliability and low-latency need to be satisfied simultaneously for short-packet transmissions [4]–[6].

A. Background

Emerging URLLC applications include virtual/augmented reality, public safety, factory automation, and autonomous vehicles, among others [7]. 3GPP has already identified several scenarios for factory automation, where actuation over industrial devices have stringent performance requirements (latency of 1 millisecond (ms) and reliability of 99.9999% [8]). Earlier studies on the long-term evolution (LTE) systems to support URLLC [9] reported that LTE is not applicable to meet low-latency requirements. The reason being its typical uplink radio access delay of 7.5 ms and handover delay of 50 ms. To solve this, a new radio interface called 5G new radio (NR) has been standardized. In 5G NR, a number of advances, including waveform numerology, frame structure, multiple access schemes, and scheduling policy, have been adopted for enabling URLLC [10]–[12].

Of particular relevance to this work is the novel grant-free (GF) random access procedure in NR. GF random access is to overcome the drawbacks of the conventional random access procedure in LTE, which is now a key enabler to support uplink URLLC. The traditional idea in LTE follows the grant-based random access procedure for a device to access channel, where it needs to request and obtain access grants via four-way handshaking with a base station (BS) [13]. Such handshaking ensures that the device has an exclusively reserved channel for contention-free transmission, avoiding any potential collisions in data transmissions. However, it incurs additional latency and undesirable signalling overheads, which hinder achieving the required level of latency constraints for URLLC. For this reason, the request-grant handshaking procedure has been removed in GF to attain prompt channel access (no sending request and waiting time for channel grant [10]).

In GF URLLC, since multiple competing devices can transmit data over the same channel, it results in potential transmission collision, which is detrimental to the reliability. To alleviate the issue, various hybrid automatic repeat request (HARQ) retransmission schemes have been considered in GF URLLC in academic research and standardization bodies [12],

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) funded by the Korea Government through the Ministry of Science, ICT and Future Planning (MSIT) through the Development of Beyond 5G Mobile Communication Technologies (Ultra-Reliable, Low-Latency, and Massive Connectivity) and Combined Access Technologies for Cellular Based Industrial Automation Systems under Grant 2017-0-00724.

Jie Ding (*Corresponding author*), Mahyar Nemati, Shiva Raj Pokhrel, and Jinho Choi are with the School of Information Technology, Deakin University, Geelong, VIC 3220, Australia (e-mail: yxdj2010@gmail.com, {nematim, shiva.pokhrel, jinho.choi}@deakin.edu.au).

Ok-Sun Park is with the Electronics and Telecommunications Research Institute, Daejeon, South Korea (e-mail: ospark@etri.re.kr).

Fumiyuki Adachi is with the Research Organization of Electrical Communication, Tohoku University, 980-8577 Japan (e-mail: adachi@ecei.tohoku.ac.jp).

[14], [15]. For instance, except traditional reactive scheme, the state-of-the-art HARQ retransmission schemes include the K-repetition scheme, where devices blindly retransmit data packets multiple times before receiving feedback, and the proactive scheme, where devices proactively retransmit until receiving positive feedback. In addition, transmission schemes such as multi-connectivity [16] and fast retrial [17] have been proposed to support GF URLLC. Nevertheless, with limited latency budget and wireless resources, their resulting reliability levels and spectral efficiency still needs to be enhanced to meet what is required for emerging URLLC services, particularly when the URLLC access load is relatively high [6].

In recent years, real momentum has been building up for deploying and advancing multiple-input multiple-output (MIMO) to enhance wireless communications' reliability and spectral efficiency [18]–[20]. Massive MIMO (mMIMO) [21], conceived in 2010, has the capability to handle a large number of antennas at a BS to serve IoT devices simultaneously. It has now been widely investigated and helped enhance several areas of wireless communications [22]. Thanks to mMIMO advances, large spatial diversity and multiplexing gains coupled with a large array gain can now be realized seamlessly. In addition, such advancements of mMIMO can be exploited to significantly enhance transmission reliability and spectral efficiency along with other critical performance metrics [23]–[28]. Thus, mMIMO has the potential for a substantial impact on GF random access in supporting a large number of URLLC devices [29]–[31].

Looking further ahead, a new distributed mMIMO architecture, so-called cell-free (CF) mMIMO, has now attracted a lot of attention from researchers and industry verticals [32]–[34]. CF mMIMO can be one of the key enablers for 6G wireless networks [35]. In a sharp contrast with (centralized) mMIMO (where co-located antennas are deployed in a compact area), antennas are spread out in CF mMIMO over a large area to serve devices without the notion of cell boundaries [32]. Since CF mMIMO reaps all the benefits from mMIMO as well as network MIMO [36], we anticipate that CF mMIMO will open up new avenues and advances for GF URLLC.

In the literature, there are several survey and tutorial papers on URLLC and mMIMO technology. For instance, a comprehensive overview of the physical layer design in NR for URLLC, including fundamental changes compared to LTE systems, can be found in [37]. The building principles of URLLC at the high layer were highlighted in [38]. A review of challenges and approaches for enabling URLLC within mMTC was provided in [3]. In [4], the principles of access protocols for URLLC and the fundamental tradeoffs from a communication-theoretic perspective were discussed. In [5], the authors focused on the study of URLLC at the network layer, where some of the key enablers of URLLC were introduced, and various mathematical tools tailored to the unique features of URLLC were examined. In [39], a tutorial on how to combine theoretical knowledge and deep learning algorithms to optimize URLLC in a cross-layer manner was investigated. In addition, opportunities and challenges in exploiting mMIMO for IoT connectivity were identified in [2] [31] and review studies on CF mMIMO can be found in [33] [34]. Unlike the existing articles, this survey mainly focuses on enabling GF URLLC and addressing its fundamental issues by exploiting mMIMO from the physical and data-link layers perspective. To the best of our knowledge, there still lacks a comprehensive investigation and insight into the state-of-the-art and potential opportunities for enabling GF URLLC in mMIMO/CF mMIMO.

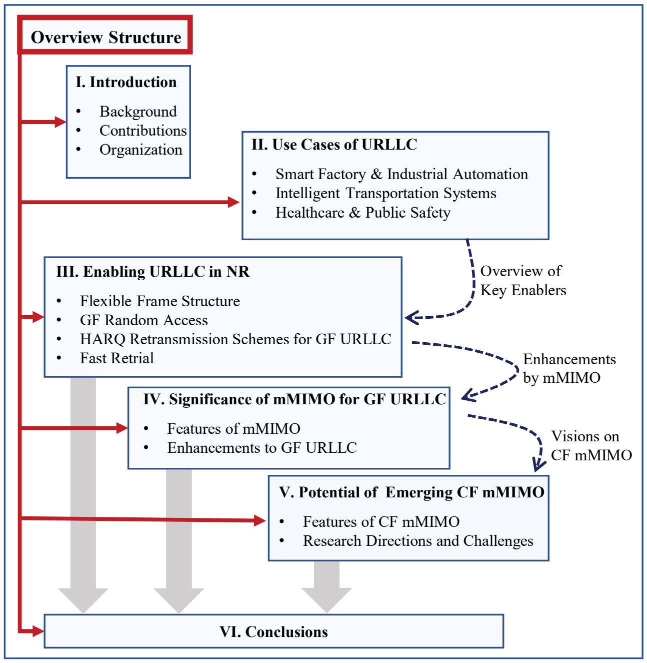


Fig. 1. Organization of the paper and interactions among different sections.

B. Contributions

As GF URLLC and its enhancements by mMIMO/CF mMIMO have been attracting more and more attention from academia and industries, it is worth aggregating the existing findings and laying a foundation for future research direction on these subjects. With this as a primary objective, our three main contributions in this paper are outlined as follows:

1) We develop a comprehensive review of NR specifications and techniques for URLLC, discussing underlying principles and highlighting fundamental issues of enabling URLLC with GF random access.
2) We review the key phenomena of mMIMO and build several deep insights into how mMIMO can be exploited to address the fundamental issues and enhance the performance of GF URLLC. In particular, we explain the benefits of exploiting preamble-collision information, coded random access, and multi-preamble detection for GF URLLC, which are only achievable via mMIMO.
3) We consider the potential of CF mMIMO and examine its distinctive features and benefits over centralized mMIMO to address impeding GF URLLC bottlenecks. Based on our understanding, we project future research directions and challenges of enabling GF URLLC in CF mMIMO.

C. Organization

A high-level simplified view of the paper structure is shown in Fig. 1. In Section II, we provide an overview of typical 5G URLLC use cases, highlighting the benefits of adopting GF random access and mMIMO to enable these use cases. To better understand the main enablers of URLLC, Section III provides a comprehensive review of the principle of key enablers in 5G NR, including GF random access, and identifies fundamental issues of enabling GF URLLC. Consequently, in Section IV, we do an in-depth investigation on the significance of mMIMO for GF URLLC and how the combination of these techniques is applied to help address the identified issues. Furthermore, Section V provides insights into future research directions focusing on the potentials of CF mMIMO for GF URLLC. Lastly, Section VI concludes the paper.

II. Use Cases of URLLC

URLLC use cases set stringent transmission requirements, e.g., 99.999% transmission success rate within 1 ms [40]. We herein focus on the three main real-life 5G business use cases: i) Industry 4.0 smart manufacturing [41], ii) Connected autonomous vehicles [42], and iii) robotic surgeries [43], all of which are now in the deployment phase. Since GF random access has been a part of 5G standards, it is worth noting that low delay and time-sensitive network applications (e.g., automated car driving and near real-time robotics) of the three use cases can benefit significantly with the GF URLLC [44]. Refer to [3], [7], [40], [44]–[48] and references therein to see other relevant URLLC use cases in different industry verticals.

Significant trends in such applications are user-specific three-dimensional (3D) video rendering, augmented reality, remote control (e.g. remote robotics, surgery, tactile internet, etc.), wireless communication automation for efficient production facilities, vehicular traffic efficiency/safety, and mobile gaming, among others. As shown in Fig. 2, we briefly review three applications of URLLC along with their requirements in the following.

A. Smart Factory and Industrial Automation

Smart factory as part of industrial automation is one of the key applications demanding URLLC. This is reflected, for example, in the Industry 4.0 migration, the process control mechanisms have been automated and deployed using the advances in wireless networks [41]. Indeed, Industry 4.0 deployment includes different ranges of real-time interface and actuation such as: process automation, motion control, industrial Ethernet, power system automation, and other control-cum-communication requirements [47], [49]. To enable factory automation, it requires real-time interactions among multiple machines, devices, robots and plants. In addition, it is worth noting that missing a deadline can be very costly as the automation systems are always cascaded and characterized by the need to meet their inter- and intra-system deadlines [50].

Furthermore, ultra-tight synchronization should be considered to be the third axis of URLLC when targeting such critical use cases for industrial automation [46]. For instance, synchronization accuracy of about 5μs is required for fault location identification [7]. The latency requirement becomes even tighter in automation processing monitoring and motion control with 50 ms and 1 ms along with the reliability of 99.9% and 99.9999%, respectively [45].

Recently, 5G NR has been considered an appropriate infrastructure for real-time interactions among such industrial entities. In [51], GF random access in 5G NR with priority-based schemes was studied in an industry automation scenario. It is shown that the GF schemes outperform legacy ones in supporting Industry 4.0 requirements. In addition, [52] studied the performance of GF URLLC in a factories-of-the-future scenario and indicated that using spatial diversity techniques appears to be the most suited strategy to achieve ultra-reliable GF transmission within a target latency.

B. Intelligent Transportation Systems

Intelligent transportation systems (ITS) is another important application of URLLC, which enables communications between transportation entities, including different types of vehicles, unmanned aerial vehicles (UAVs), trains, roadside units etc. One prevalent example is the Internet of Vehicles (IoV), wherein on-vehicle distributed learning models are trained by exchanging inputs, outputs, and their learning parameters in an ad hoc fashion [53]. To be specific, the latency requirement for such distributed learning and connections of the ITS entities with the infrastructure backhaul should be lower than 10 ms, along with a reliability of 99.9999% [45]. For example, missing deadlines in applications of autonomous vehicles can be highly catastrophic [42]. Also, we are still in an early phase of ITS deployment, and we have been suffering from limitations in case of emergencies which results in larger delays in broadcasting emergency messages [54], [55]. Such dissemination demands URLLC and becomes more critical in the case of high-speed trains and aerial vehicles [56]. Besides, for vehicle to vehicle (V2V) communication in connected autonomous vehicles, the latency requirement of 1 ms is mandatory. Simultaneously, with the current IEEE 802.11p protocol, V2V communication suffers from unbounded latency and highly varying reliability [53], [57].

Motivated by these, GF random access has been considered and identified as a key enabler for 5G empowered ITS [44]. Furthermore, as indicated in [58], the emerging mMIMO technology can facilitate gigabits-per-second (Gbps) communication for a variety of cellular ITS scenarios, which is far beyond what the legacy dedicated short-range communication (DSRC) [59] and LTE-advanced systems can support.

C. Healthcare and Public Safety

Health care applications such as robotic telesurgery and tactile Internet spur the need for URLLC. The remote surgical consultations and remote surgery, so-called *telesurgery*, adopt augmented reality (AR) and virtual reality (VR) advancements. For instance, surgeons use VR headsets to observe/actuate the inside of a patient. They also require patients over VR headsets when taking them through their surgical plan [47]. The connectivity requirements for the explosive growth of such devices and systems with sensor-based implementations in healthcare

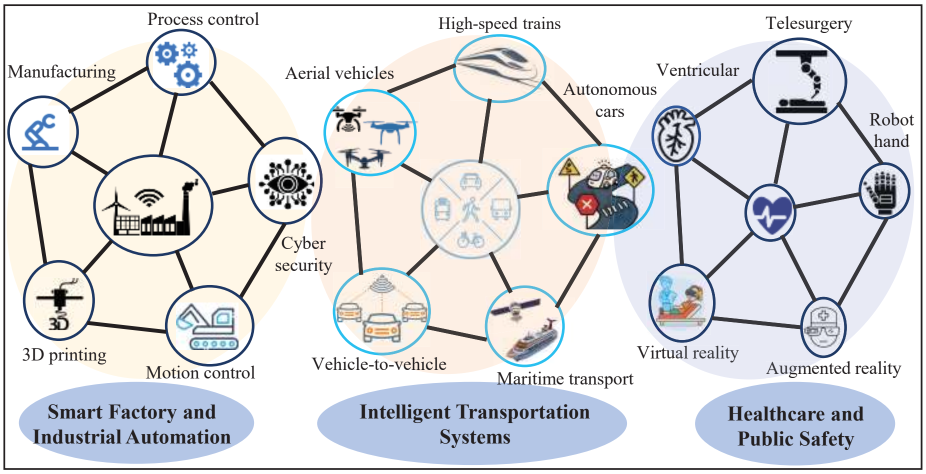


Fig. 2. Three real life use cases of URLLC. Among others, the significance of URLLC in these use cases can be observed in terms of the following three examples: i) aids in the automation of manufacturing operations and factory control systems; ii) facilitates surgical augmented-reality which allows surgeons to undertake surgical resection in a much more precise and analytical manner, reducing the risk of relapse; iii) enables route discovery and collision avoidance in real-time relying heavily on communications involving cars and everything else in the transportation grid.

facilities will accelerate the growth of mMTC. The required end-to-end latency for such application of health care verticals are low, ranging from 125 ms to 1 ms (for mission-critical healthcare applications) with 99.9999% reliability guarantees [43].

Likewise, public safety requires robust and reliable communications in case of natural disasters such as earthquakes, tsunamis, floods and hurricanes [60]. Specifically, accurate positioning, fast communications, real-time video, and the ability to send high-quality pictures in the presence of damaged wired technologies are critical challenges for public safety [49].

Various studies have been carried out to enable applications in both domains. For instance, [61] advocated the use of GF and non-orthogonal transmissions to achieve a significant reliability gain for tactile Internet. And [62] reviewed emerging technologies for enabling public safety services, and mMIMO is one of them.

Overall, emerging URLLC use cases are posing unprecedented challenges in terms of latency, reliability, and scalability for systematic design [63]. As reviewed, GF random access and mMIMO have the potential to address the challenges [31], [44]. To provide a comprehensive and insightful view, in the following, we first present the principles of enabling URLLC in 5G NR, including GF random access and identify fundamental issues of enabling GF URLLC. Based on them, we then shed light on the potential of ameliorating the issues and advancing URLLC capabilities with the assistance of mMIMO technology, which can be served as guidelines for underpinning the applications of URLLC.

III. Enabling URLLC in NR

To meet latency and reliability requirements of various URLLC use cases, key enabling techniques in NR physical and medium access control (PHY/MAC) layers have been adopted in 3GPP release 15 and enhanced in 3GPP releases 16 and 17 [11], [49], [64]. In this section, we explain several key NR specifications and techniques for URLLC, including flexible frame structure, GF random access, and retransmission schemes.

A. Flexible Frame Structure

In NR, one of the main specifications adopted for URLLC is a flexible frame structure, which is able to not only reduce latency but also create more retransmission opportunities within a target latency that in turn lead to enhanced reliability [45].

Since the user-plane, latency[1] is one of the dominant components in the latency of URLLC and the transmission time interval (TTI) plays an important role in contributing to the user plane latency, reducing TTI is a key to meet the low-latency requirement. For this reason, NR enables TTI reduction by introducing scalable numerology (subcarrier spacing) and the concept of mini-slots [49].

Specifically, the subcarrier spacing in LTE is fixed at 15 KHz, and the basic TTI is set to 1 ms, which equals the length of a subframe/slot. Different from it, the TTI of URLLC can be shortened by increasing the subcarrier spacing and/or using small scheduling units such as mini-slots. As illustrated in Fig. 3, the subcarrier spacing of $2^n \times 15$ KHz can be configured for URLLC data transmissions depending on bandwidth, deployment, and use cases, where n can be $0, 1$, or 2 in Sub-6 GHz and 3 in the millimeter-wave spectrum. Since a larger subcarrier spacing can be employed than that of the baseline of 15 KHz, the duration of slot and orthogonal frequency division multiplexing (OFDM) symbol can be significantly reduced. For instance, the slot-based TTI with 60 kHz subcarrier spacing is 0.25 ms, which is only a quarter of that with 15 KHz.

On top of this, the number of OFDM symbols in each TTI does not necessarily equal 14. To be more precise, a mini-slot can be employed to shorten further the TTI, consisting of 1 to 13 OFDM symbols. In Fig. 3, a mini-slot based TTI of 2

[1] User-plane latency is the time it takes to successfully deliver a data packet at the radio protocol layer from the transmitter to the receiver.

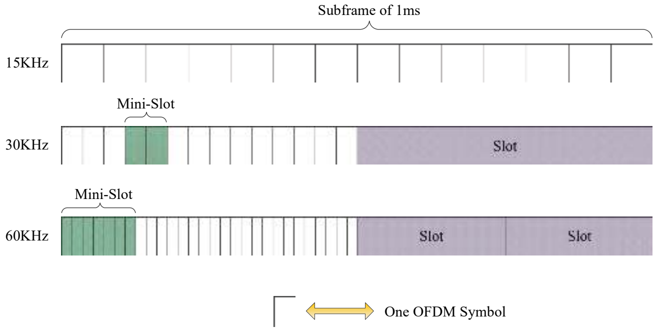


Fig. 3. Illustration of slot and mini-slot structure for different numerologies [49].

OFDM symbols with the subcarrier spacing of 30 KHz and a mini-slot based TTI of 7 OFDM symbols with 60 KHz are illustrated, respectively. Evidently, with a subcarrier spacing of 30 KHz, a 70 μs TTI is achieved by a 2-symbol mini-slot as opposed to 0.5 ms based on slot-based transmission. This demonstrates a substantial latency reduction by adopting short TTI. Furthermore, suppose a target latency of 1 ms, when the mini-slot is adopted, 14 transmission opportunities are created within the latency, which can be exploited to enhance the transmission reliability.

B. GF Random Access

To implement a radio access network for URLLC, handling random access design can be one of the most important/critical portions. GF random access (also known as 2-step random access the [65]) procedure is another key NR specification to enable URLLC.

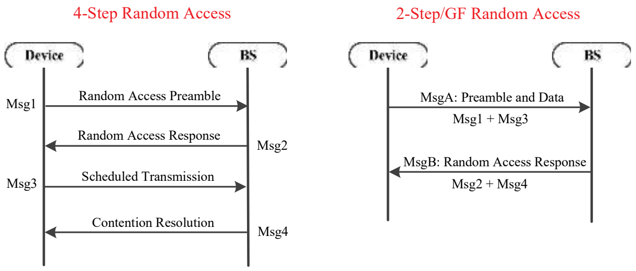


Fig. 4. Illustration of GF random access procedure compared to conventional 4-step random access procedure [11].

In LTE, a 4-step random access procedure, also known as grant-based random access, is adopted, which is mainly designed for human-type communication and not suitable for MTC, including URLLC [9], [66]–[68]. As illustrated in Fig. 4, in the first step of the 4-step random access, each active device transmits a randomly selected preamble on physical random access channel (PRACH) to initiate a scheduling request. In the second step, the BS detects the preambles transmitted by active devices and sends responses by issuing a scheduling grant. Once an active device is connected to the BS, and it can transmit data packets in the third step on dedicated resource blocks (RBs) or channels, which are physical uplink shared channel (PUSCH). The 4-step random-access procedure requires two round-trip cycles between the devices and the BS, which raises the barriers to meet the stringent latency requirement of URLLC use cases [9]. In particular, it not only increases the latency but also incurs large control-signaling overhead for small packets [11].

The motivation of GF random access is to reduce latency and control-signaling overhead by having a single round-trip cycle between the devices and the BS. Compared to the 4-step random access, GF random access can be more efficient thanks to low signaling overhead when devices have short packets to transmit. In GF random access, an active device does not wait for a scheduling grant from the BS. That is, once a device becomes active in a random-access slot (TTI), it is to transmit a preamble directly along with data on the same channel in a time division multiplexing (TDM) manner and waits for the acknowledgement from the BS.

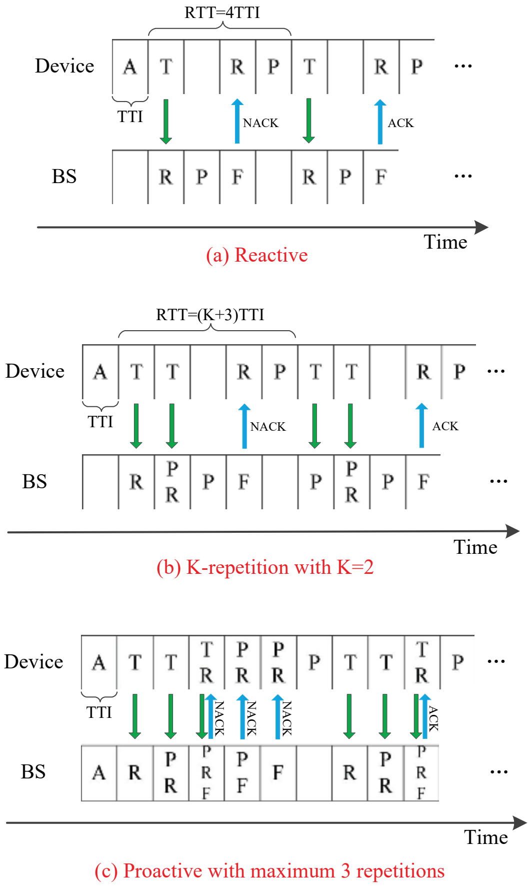
In GF random access, contention-free and contention-based transmission modes can be operated. In this paper, we mainly focus on the latter transmission mode. The contention-free GF transmission can be employed when the number of active devices is small and their access traffic is periodic or deterministic. It allows the BS to pre-allocate wireless resources, e.g., preamble and spectrum resources, to devices so that access contention can be avoided [69], [70]. In the case of sporadic traffic or/and massive URLLC, contention-based GF transmission is more suitable since it is more efficient and flexible in terms of resource utilization [71]. Nevertheless, contention-based GF transmission is prone to potential access collisions when multiple devices simultaneously access the same channel resource, thus jeopardizing transmission reliability [52], [71], [72].

C. HARQ Retransmission Schemes for GF URLLC

There are a number of techniques to improve transmission reliability. We briefly discuss HARQ retransmission schemes for URLLC in this subsection. Thanks to the flexible frame structure in NR, as mentioned above, multiple retransmission opportunities can be created within a target latency by adopting short TTI. As a result, several different HARQ retransmission schemes can be incorporated into GF random access to support URLLC.

1) Reactive HARQ Retransmission Scheme: Reactive HARQ retransmission scheme is a conventional scheme adopted in LTE. In this scheme, retransmission is allowed only when a device receives a Negative ACKnowledgement (NACK). Once the device transmits a data packet, the device has to wait for feedback from a BS before any retransmission attempt. To issue the feedback, the BS is to receive and process the data packet from the device. In particular, as shown in Fig. 5(a), with the assumption that both the BS and the device spend 1 TTI for processing and 1 TTI for transmitting [74], we see that the HARQ round-trip-time (RTT[2]) takes 4 TTI in the scheme, which means that the device needs to wait for 4 TTI until the next retransmission attempt. In this case, if the

[2] RTT is the time duration of the cycle from the beginning of transmission until processing its received feedback [73].



(a) Reactive

(b) K-repetition with K=2

(c) Proactive with maximum 3 repetitions

Fig. 5. Illustration of HARQ retransmission schemes in GF random access procedure for URLLC [73], where ‘A’ represents frame alignment, ‘T’ represents transmission, ‘R’ represents reception, ‘P’ represents processing, and ‘F‘ represents feedback.

maximum URLLC latency is up to 10 TTI, each device can only have one retransmission.

2) K-Repetition HARQ Retransmission Scheme: To further reduce the latency and improve the reliability, the K-repetition HARQ retransmission scheme has been proposed and adopted in 3GPP release 15 [15], [75]. As shown in Fig. 5(b), each device transmits the same data packet K times before receiving feedback from the BS, and the BS can perform a soft combining of these packets to enhance the reliability. In this scheme, each data repetition can be identical or be a different redundancy version of the encoded data. Clearly, compared to the reactive scheme, this scheme has the potential to reduce the latency given the same number of retransmissions. Nevertheless, the predefined value of K needs to be carefully selected since the time and resources would be wasted if it is overestimated [73].

3) Proactive HARQ Retransmission Scheme: In addition, a proactive version of the K-repetition scheme, namely the proactive HARQ retransmission scheme, was studied in [14], [76]. It is also known as repetitions with early termination. As shown in Fig. 5(c), in this scheme, each device consecutively transmits the same data packet with a maximum of K times, but it will stop transmitting if it receives an ACK from the BS in the middle of the transmissions. Compared to the K-repetition scheme, this scheme incurs additional computational complexity and energy consumption for devices since they need to monitor and process the feedback more frequently. However, it is likely to be more resource-efficient if the value of K is overestimated [73]. It is important to note that although the proactive scheme is implementable in GF random access, it is not a part of the 3GPP standard.

Several research works have provided comprehensive studies on the performance of the three HARQ retransmission schemes in GF URLLC random access [73], [77], [78]. In [73], the authors evaluated the performance in a large outdoor urban macro scenario using an extensive system level simulations and showed the superiority of GF transmission to conventional grant-based one. In a [77], the authors not only presented an overview of the three schemes in GF URLLC, but also introduced non-orthogonal multiple access (NOMA) based HARQ retransmission schemes for beyond the 5G standard. The NOMA based schemes are similar to the coded random access [79] scheme that employs successive interference cancellation (SIC) at the receiver to remove the interference from decoded data packets. In [78], the authors presented a Spatio-temporal analytical framework by considering the impact of preamble collision for contention-based GF random access. Based on this framework, they defined the latent access failure probability to characterize URLLC reliability and latency performance. A tractable approach was also proposed to derive and analyze the latent access failure probabilities of a device with the three different schemes. Based on the analysis and results of these works, the following observations can be made.

- The GF random access with HARQ retransmission schemes achieves lower latency than the grant-based counterpart under given target reliability.
- Although the reactive scheme brings a longer latency tail, it is more suitable to accommodate moderate URLLC load. The K-repetition scheme has the potential to achieve the lowest latency under a light URLLC access load. Nevertheless, it is difficult to fulfill the URLLC requirements when the access load increases. For the proactive scheme, its performance is similar to that of the K-repetition scheme when $K \leq 4$, but with higher spectral efficiency when $K > 4$.

It is important to remark that the performance limitations for the HARQ retransmission schemes in GF random access are fundamentally originated from the multi-user interference as well as preamble collision when multiple devices compete for the same channel resource for uplink transmissions. Taking the K-repetition scheme as an example, although its multiple consecutive transmissions can increase the combining gain, it also introduces additional multi-user interference and potential preamble collision, which could surpass the benefits of the combining gain and result in a low spectral efficiency in conventional systems, making it unable to accommodate high URLLC load.

In fact, in GF random-access, the BS has to identify a device and estimate its channel state information (CSI) by detecting the preamble before attempting to recover any data using coherent decoding. Thus, when a preamble collision occurs, i.e., multiple devices select the same preamble over the same channel resource, the BS is unable to estimate their CSI, which means that data packets cannot be decoded. Unfortunately, most HARQ schemes for URLLC overlook the importance of CSI estimation. For example, the analysis approach in [78] does not consider the impact of CSI estimation error on the performance, which makes the analysis result too optimistic.

D. Fast Retrial

While HARQ is a link-layer protocol that uses retransmissions for reliable transmissions, different retransmission schemes are developed in random access. In particular, in random access, retransmissions are required as packet collisions are inevitable due to the nature of uncoordinated transmissions. To avoid consecutive packet collisions, various backoff algorithms with random backoff times can be employed for random access [80]. Provided that there are multiple channels (e.g., multichannel ALOHA), it is possible for devices to immediately retransmit their data packets without random backoff times, which is referred to as fast retrial [81]. Fast retrial can reduce the access delay and, as a result, it has the potential to meet URLLC requirements. In [17], fast retrial is employed for GF random access, and it is shown that the access delay can be shortened. In [82], repetition diversity is considered to take advantage of retransmissions by fast trial in GF random access with mMIMO.

It is noteworthy that most HARQ schemes for URLLC are studied as a link-layer protocol for a single device under the assumption that the signals from the other competing devices are considered multi-user interference. While this can simplify the analysis, it does not allow seeing the overall performance as a multiuser or random access system and to take into account any impact of preamble collision and CSI estimation error on the performance as mentioned earlier.

Consequently, it is necessary to study retransmission schemes within a GF random access system where multiple devices are competing for the shared resource to meet URLLC requirements. In particular, it is expected that the fundamental issues such as preamble collision and resulting CSI estimation error can be effectively mitigated or addressed so as to achieve the desired URLLC capability in GF random access. To this end, mMIMO can be exploited.

IV. Significance of mMIMO for GF URLLC

To improve the performance of GF URLLC, mMIMO can be considered, which has been identified as one of the essential technologies in 5G and towards 6G [83]–[85]. This section briefly presents the key features of mMIMO and discusses what they can provide to GF URLLC for better performance.

A. Features of mMIMO

In 3GPP release 15, key aspects of mMIMO have been included [83]. In particular, a maximum of 256 antennas can be supported, compared to the 64 antennas in release 13. Thanks to the excessive degrees of freedom created by a large-size antenna array, two prominent features such as channel hardening and favorable propagation can be achieved in mMIMO [23], [86], [87].

Channel hardening means that the effect of small-scale fading is averaged out, and devices' channels behave a deterministic like wired channel as the number of antennas approaches infinity [86]. Given that the CSI of an arbitrary device (taking device 1 as the device of interest in Fig. 6) is denoted by $\mathbf{h}_1$, it can be described by a random vector with each element distributed as $\mathcal{CN}(0,1)$ when an independent Rayleigh fading channel model is considered. Letting M denote the number of antennas, based on the law of large numbers, we have $\frac{\|\mathbf{h}_1\|^2}{M} \xrightarrow{M\to\infty} 1$ and $\mathrm{Var}\left\{\frac{\|\mathbf{h}_1\|^2}{M}\right\} = \frac{1}{M}$.

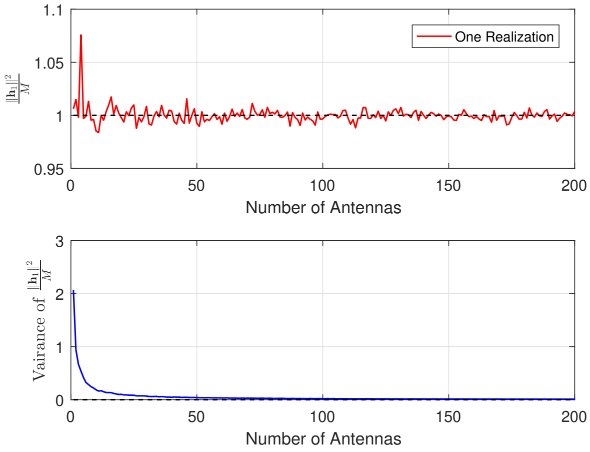


Fig. 6. Illustration of channel hardening as the number of antennas, M, increases for independent Rayleigh fading channels.

Fig. 6 presents an illustration of the channel hardening feature as a function of M. It is seen that the variance of $\frac{\|\mathbf{h}_1\|^2}{M}$ decays with M and converges towards zero, which indicates that the channel gets more hardened as M increases.

Favorable propagation means that channels of different devices become orthogonal as M approaches infinity [87], which makes different devices distinguishable in the space domain. Considering two different devices (device 1 and device 2) and by using the law of large numbers, we have $\frac{\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_2}{M} \xrightarrow{M\to\infty} 0$ and $\mathrm{Var}\left\{\frac{\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_2}{M}\right\} = \frac{1}{M}$.

Fig. 7 presents an illustration of the favorable propagation feature as a function of M. It is seen that the channels of different devices become orthogonal, and the variance of $\frac{\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_2}{M}$ converges to zero as M gets larger.

B. Enhancements to GF URLLC

By taking advantage of these two features, mMIMO is able to spatially separate signals that are simultaneously transmitted by a large number of URLLC devices over the same channel resource in GF random access, which makes it a prominent enabler for massive access [31]. Furthermore, simple linear processing such as conjugate beamforming and zero-forcing (ZF) beamforming can achieve near-optimal performance with favorable propagation [88]. In this subsection, we present a number of potential enhancements to GF URLLC by mMIMO.

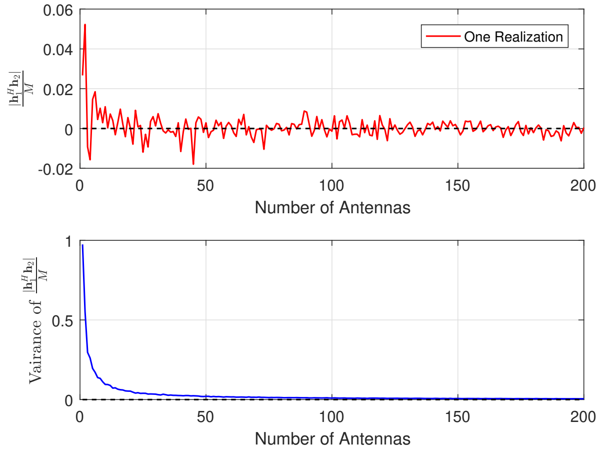


Fig. 7. Illustration of favorable propagation as the number of antennas, M, increases for independent Rayleigh fading channels.

1) High Reliability and Capacity: One of the most significant enhancements to GF URLLC is the high reliability and capacity ensured by mMIMO, thanks to its large array, diversity, and multiplexing gains. In GF random access with high access loads, several URLLC devices can become active simultaneously and share the same TTI for data transmissions/retransmissions. In conventional systems with BSs equipped with few antennas, even without preamble collision, it is still highly likely that their data decoding can fail. Their reliability can hardly meet requirements due to the impact of multi-user interference and noise [73], which inevitably triggers extra retransmission attempts or may lead to access failure, increasing the latency and degrading the transmission reliability as a result. In mMIMO, as explained in Section IV-A, the channel of a device can behave like wired channels and multi-user interference as well as noise will vanish as M approaches infinity [89]. Therefore, high transmission reliability and system capacity can be achieved to serve a large number of URLLC devices while coping with stringent requirements.

To provide a glimpse into the advantage of employing mMIMO for GF URLLC, the decoding error probability in the finite blocklength regime [90] is considered. Based on [14], [91], [92], the decoding error probability of receiving q bits of data within d channel uses can be well approximated by

$$\epsilon(\gamma) \approx \mathbb{E}_{\{\gamma\}}\left[\mathcal{Q}\left(\frac{d\log_2(1+\gamma)-q}{\sqrt{V(\gamma)d}}\right)\right], \qquad (1)$$

where $\mathbb{E}_{\{x\}}[\cdot]$ is the expectation operation over variable x, $V(\gamma)$ is the channel dispersion that is given by $V(\gamma) = \left(1-\frac{1}{(1+\gamma)^2}\right)\log_2^2(e)$. Here, γ denotes the instantaneous signal-to-interference-plus-noise ratio (SINR) and $\mathcal{Q}(x)$ is the Q-function.

In a multi-user mMIMO scenario with N URLLC devices and M antennas, the instantaneous SINR of device n can be given by [24]

$$\gamma_n = \frac{\rho|\mathbf{b}_n^{\mathrm{T}}\mathbf{h}_n|^2}{\rho\sum\limits_{\substack{i=1\\ i\neq n}}^{N}|\mathbf{b}_n^{\mathrm{T}}\mathbf{h}_i|^2+\|\mathbf{b}_n\|^2}, \qquad (2)$$

where ρ denotes the received signal-to-noise ratio (SNR) per antenna and $\mathbf{b}$ represents the receive beamformer. Under an ideal assumption that the CSI of devices is perfectly obtained by the BS, we have $\mathbf{b} = \mathbf{h}^*$ in the case of conjugate beamforming. When $M \to \infty$, asymptotic analysis can be used. Based on the law of large numbers, the asymptotic form of (2) can be written as

$$\tilde{\gamma}_n \approx \frac{\rho M}{\rho(N-1)+1}. \qquad (3)$$

As we can see from (3), with a fixed N, a larger M leads to an increase of γ_n and, as a result, improved reliability. Furthermore, for a given required SINR γ, more devices can be served simultaneously with a larger M.

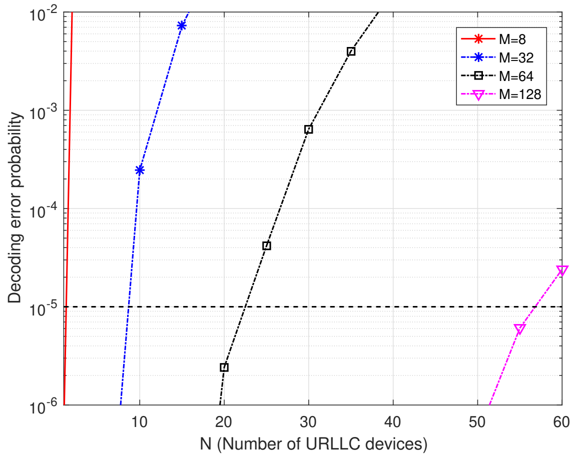


Fig. 8. Decoding error probability as a function of N. In this simulation, we set that $d = 200$ with bandwidth of 200 KHz and latency of 1 ms, $q = 200$, and $\rho = 0$ dB.

With (1) and (2), the decoding error probability as a function of N with different values of M is plotted in Fig. 8, where an independent Rayleigh fading channel model is assumed. As observed, with a fixed N, the decoding error probability can be significantly reduced as M grows, which reveals that the need for retransmissions can be dramatically reduced in GF URLLC. Furthermore, with target reliability of $\epsilon = 10^{-5}$, it is evident that only 1 URLLC device can be served in MIMO with 8 antennas, while the number of URLLC devices that can be served simultaneously increases with M and can reach 55 when $M = 128$. These exemplary results confirm the advantage and potential of mMIMO to support GF URLLC in the case of high access loads. It is noted that more devices can be supported, and the reliability can be further improved when ZF or minimum mean-square-error (MMSE) beamforming [93] is used in mMIMO [30], [88].

It is also important to remark that accurate acquisition of instantaneous CSI is a key to fully exploiting large spatial diversity and multiplexing gains of mMIMO. The above exemplary results can be optimistic since perfect CSI is assumed to be known at the BS. In practice, it is not directly available at the BS in the context of GF random access. As introduced in Section III-B and III-C, the CSI of devices in GF random access is estimated by detecting devices' selected preambles. Since preamble resources are finite [94], there is a non-zero probability that multiple active devices select the same preamble, which inevitably leads to preamble collision (especially in the case of high URLLC load). Since the estimated CSI is a noisy superposition of multiple channel vectors of the collided devices, it brings two effects in mMIMO, leading to unsuccessful decoding [89], [94]–[96]: 1) it reduces the coherent array gain of the desired received signal, and 2) it introduces a coherent interference that gets stronger as M grows. Since preamble collision in GF URLLC is dominantly detrimental to the transmission reliability, leading to increased latency and need for retransmissions, it is desirable to effectively mitigate it. In the following, enhancements in this respect by exploiting mMIMO are presented.

2) Utilization of Preamble-Collision Information: In the context of preamble collision, a distinctive phenomenon introduced by mMIMO is that preamble-collision information can be made available at the BS, which exceeds the capability of conventional systems [97]. In particular, the preamble-collision information can include the preamble multiplicity, i.e., the number of devices selecting the same preamble.

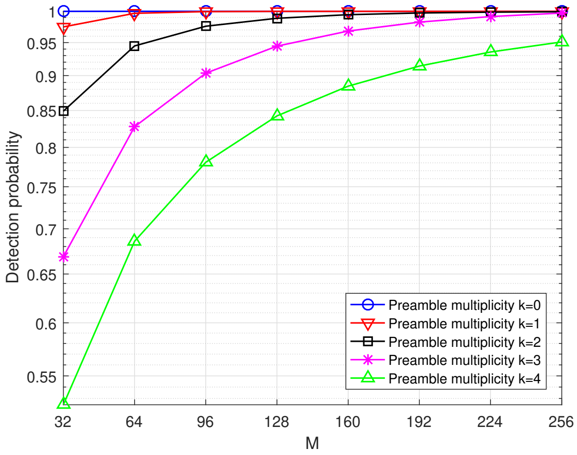


Fig. 9. Detection probability of preamble multiplicity as a function of M with different preamble multiplicity k, where received SNR ρ=0 dB and the number of preambles is set to 25 [98].

Several research works, e.g., [98]–[100], have analytically demonstrated that accurate preamble multiplicity detection can be achieved using a simple energy detector in mMIMO. In Fig. 9, detection probability of preamble multiplicity is illustrated with respect to M. It is evident that the detection probabilities with different preamble multiplicity k approach 1 as M increases. In addition, the detection accuracy is degraded as the preamble multiplicity increases. This is because a larger preamble multiplicity leads to a larger variance of the received preamble-signal energy. Therefore, it is comparatively easier to get confused with adjacent multiplicities [98]. Nevertheless, we still see that even for a large preamble multiplicity of $k = 4$, the high detection accuracy of 84.7% can be achieved at $M = 128$.

The preamble-collision information is particularly beneficial that allows the BS to identify the number of collided/uncollided devices, and determine the URLLC load of GF random access as well as trends regarding its changes. As a result, when it is available at the BS, it could facilitate, e.g., reengineering of GF random access procedure to alleviate and control preamble collision [77], [82], [101] or adaptive resource partitioning for URLLC devices to improve the transmission reliability and throughput [98], [102]. In [98], for instance, by exploiting the preamble-collision information, dynamic preamble-resource partitioning (DPP) schemes in GF URLLC with reactive retransmission scheme were developed to reduce the access failure rate under a given latency constraint. In the DPP schemes, with the estimated preamble multiplicities, the number of URLLC devices with preamble collision that will retransmit in the next transmission slot is inferred at the BS. Based on the inferred information, the BS can strategically divide the whole preamble resources into two separate sets to accommodate different groups of devices, i.e., retransmission devices and newly arrived devices. It was shown in [98] that under given latency budget and target access failure rate, the DPP schemes with the use of preamble-collision information can increase the URLLC access load by 42% and improve the preamble-resource utilization by over 25% compared to the conventional baseline scheme without DPP.

3) Coded Random Access and Multi-Preamble Detection: In GF URLLC with HARQ retransmission schemes, devices have a chance to transmit the same packet multiple times within the target latency. For instance, the K-repetition retransmission scheme allows devices to consecutively transmit packets K times before receiving feedback. Based on this fact, the size of preamble space can be extended to L^K compared to L with a single-TTI transmission (i.e., $K = 1$), where L represents the preamble length. Since L^K exponentially increases with K and is larger than L, preamble collision can be thus significantly alleviated. In particular, with N active URLLC devices, the preamble collision probability of a device can be written as

$$P_{\mathrm{pc}} = 1 - \left(1 - \frac{1}{L^K}\right)^{N-1}. \tag{4}$$

In Table I, the preamble collision probabilities as function of K are given with $N = 10$ and $L = 48$ [78]. It is obtained that P_{pc} sharply approaches 0 as K increases, which decreases to 1.7×10^{-6} when $K = 4$ compared to 0.17 when $K = 1$. With this phenomenon, mMIMO can be utilized to resolve preamble collision, separate signals of devices effectively, and decode their data, which are beyond the capability of conventional systems.

One of the classic schemes is coded random access [79], which relies on favorable propagation and channel hardening of mMIMO. In coded random access, a frame consists of multiple time slots, and each active device transmits multiple randomly selected preambles within a frame, while the

TABLE I
PREAMBLE COLLISION PROBABILITIES WITH DIFFERENT K, WHEN $N = 10$ AND $L = 48$.

K	1	2	3	4
P_{pc}	1.7×10^{-1}	3.9×10^{-3}	8.1×10^{-5}	1.7×10^{-6}

same data packet is repeatedly transmitted. Since there can be time slots with preamble-collision-free transmissions, it allows interference-free channel estimation of some devices and then successfully decodes their data packets after receiving beamforming. The BS then applies SIC in order to remove the signals of these devices in all the slots. After SIC, the BS can further find slots with collision-free transmissions and perform channel estimation and decoding, and so on. The same process is to be repeated until the BS cannot find any slot with the collision-free transmission. In order to further improve the spectral efficiency or shorten the latency of coded random access, a superposition of multiple orthogonal preambles [103] can be considered, which can effectively increase the number of active devices whose CSI can be estimated by the BS using SIC.

Although the coded random access was originally designed for mMTC to resolve preamble collision and enhance access load, it can be extended to mMIMO straightforwardly assisted GF URLLC. In Fig. 10, a simple example of coded random access with 2-repetition retransmission scheme is illustrated. There are 3 URLLC devices that share the same channel resource, and each randomly selects preambles over two transmission TTIs from a set of preambles $\{S_1, S_2\}$. Notice that devices 1 and 3 have no preamble collision in the first and second TTIs, respectively. In such a case, the BS can first decode data of devices 1 and 3 thanks to large array gain in mMIMO. Then, SIC can be applied to remove the signals of these devices when estimating the channel and decoding data of device 2.

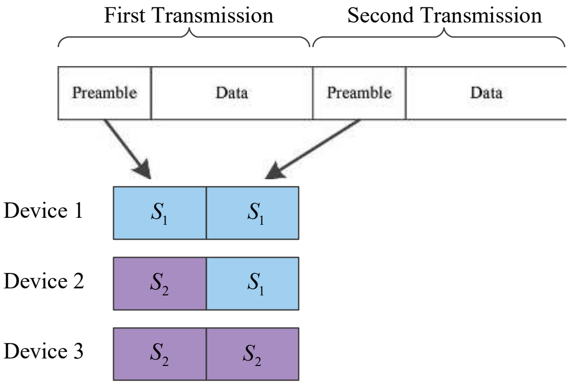


Fig. 10. A simple example of coded random access in GF URLLC with 3 active devices and 2 preambles, where S_n represents the nth preamble.

As indicated in [96] and [101], when M and L^K are sufficiently large, the transmission error probability by employing SIC in coded random access is equivalent to the preamble collision probability, i.e.,

$$P_{\mathrm{e}}^{\mathrm{SIC}} = P_{\mathrm{pc}}. \tag{5}$$

On the other hand, as shown in [104], the transmission error probability in conventional K-repetition retransmission scheme without using SIC is given by

$$P_{\mathrm{e}}^{\mathrm{Conv}} = \left(1 - \left(1 - \frac{1}{L}\right)^{N-1}\right)^{K}. \tag{6}$$

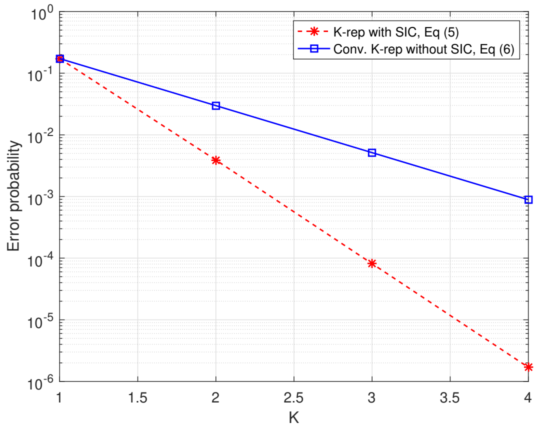


Fig. 11. Error probability comparison between the scheme with SIC and conventional scheme without SIC in K-repetition retransmission, where $N = 10$, $L = 48$, and $M = \infty$ in mMIMO.

In Fig. 11, based on (5) and (6), error probability comparison is illustrated between coded random access scheme with SIC and conventional scheme without SIC in K-repetition retransmission, where $N = 10$, $L = 48$, and $M = \infty$ in mMIMO. As observed, the employment of SIC in mMIMO for GF URLLC can significantly enhance the transmission reliability. Suppose that a URLLC application aims to achieve an error probability of 10^{-5} and its required latency only allows a maximum $K = 4$ retransmissions in GF URLLC. We see that the scheme with SIC can simply achieve the desired reliability within the required latency when 10 URLLC devices share the same time-frequency resource, while the conventional scheme without SIC is unable to.

It is important to note that in order to enable interference-free channel estimation after SIC in coded random access, the BS needs to have prior knowledge of the multi-preamble choices of devices. In [79], it assumed that the uplink data embeds the information about the random pilot and data transmission schedule of a device. Nevertheless, this incurs additional overhead and reduces the spectral efficiency. Moreover, the information can only be acquired when one of the multiple copies of data duplicates can be successfully decoded.

In fact, the multi-preamble choices of all active devices can be directly detected with received preamble signals in mMIMO without the need of being embedded into data. In [105], a low-complexity and reliable multi-preamble detection algorithm were proposed by exploiting channel hardening and favorable propagation of mMIMO. It was shown that the detection probability of a multi-preamble choice equals $1-P_{\mathrm{pc}}$

when M is sufficiently large, which indicates that the multi-preambles of a device can be accurately detected as long as it is not completely collided with that of other devices. In addition, since the detection algorithm only relies on the received preamble signals for acquiring all the multi-preamble choices, SIC is not necessarily needed.

Although mMIMO is a promising solution to enhance GF URLLC, it is a cell-centric based network infrastructure. The performance of cell-edge devices would be prone to inter-cell interference in both preamble and data domains as the network becomes densified. In addition, gathering a huge number of antennas in a centralized manner might be practically challenging when considering the array dimension and hardware cost. To overcome these issues and still reap all benefits obtained from mMIMO, CF mMIMO can be a propitious candidate [85]. In the next section, we introduce several benefits of CF mMIMO compared to mMIMO and discuss its potential research directions along with challenges for GF URLLC.

V. Potential of Emerging CF mMIMO

CF mMIMO is a new incarnation of distributed mMIMO, which is a user-centric network infrastructure [33]. In CF mMIMO, instead of gathering all the antennas at the same location, a large number of simple and low-cost access points (APs) employing single or multiple antennas are spatially distributed to serve several devices that share the same channel resource jointly. Each AP is connected to the BS central processing unit (CPU) via a fronthaul. In contrast to the concept of mMIMO, the cellular or cell boundary concepts disappear in CF mMIMO, and a device can be served simultaneously by multiple APs.

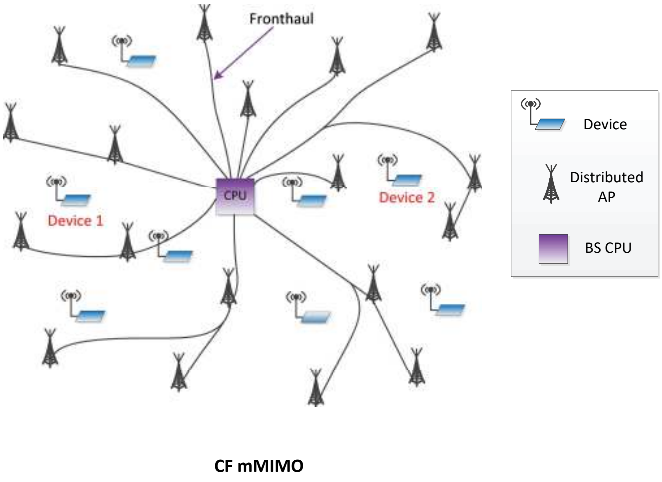


Fig. 12. An illustration of network deployment of CF mMIMO.

An example network deployment of CF mMIMO is illustrated in Fig. 12. Since CF mMIMO not only captures the benefits of mMIMO but also network MIMO [36], it provides more distinctive features than mMIMO [106]–[109]. In the following, we present its key features and benefits compared to mMIMO.

A. Features of CF mMIMO

CF mMIMO inherits the features of channel hardening and favorable propagation from mMIMO, although their level may be less than that in mMIMO (particularly channel hardening). As discussed in [110], one can expect stronger channel hardening in CF mMIMO by deploying multiple antennas per AP, and stronger favorable propagation by increasing the densities of APs and antennas per AP. Because of these facts, the enhancements to GF URLLC by mMIMO (as discussed in Section IV-B) could also be applied and extended in the context of CF mMIMO. In addition, macro-diversity and signal spatial sparsity are the most two distinctive features of CF mMIMO than mMIMO, which can be potentially utilized to further enhance the performance of URLLC in GF random access. Clearly, these two features differentiate CF mMIMO from centralized mMIMO.

1) Macro Diversity: In CF mMIMO, since APs are geographically distributed, it is highly likely that several neighbouring APs surrounds each device. As a result, the average distance from a device to any nearest AP is significantly short compared to mMIMO, which leads to the increase of macro-diversity gain.

To illustrates this feature, we present Fig. 13 with an indoor GF URLLC scenario, where URLLC devices perform GF random access to transmit data. Similar to [106], we assume a factory automation scenario with a square area of 100×100 m^2. Within this area, there are $Q = 100$ distributed APs deployed on the ceiling on a square grid. Each AP has a height of 6 m and is equipped with $S = 4$ antennas. To avoid boundary effects, the square area is wrapped around. The system works at 3.5 GHz with a bandwidth of 10 MHz, and the noise has a power spectral density of -174 dBm/Hz. Moreover, the URLLC device's transmit power is set to 23 dBm, and an additional receiver noise figure of 7 dB is considered. For the channel model, the Rayleigh small-scale fading model is assumed, and the large-scale channel model is implemented based on a 3GPP indoor industrial model, namely "clutter-embedded APs" setup [111].

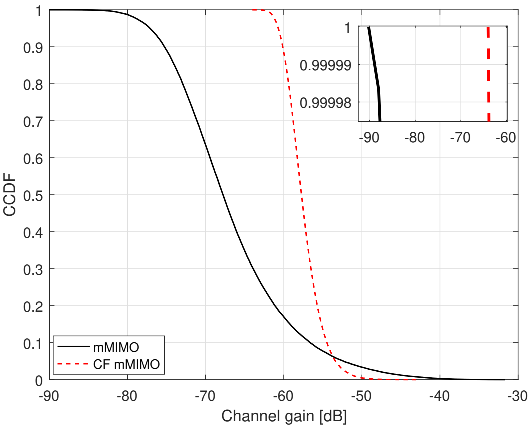


Fig. 13. Channel gain comparison between mMIMO and CF mMIMO in a factory automation scenario.

Fig. 13 shows the complementary cumulative distribution function (CCDF) of the channel gain $\|\mathbf{h}\|^2$ ($\mathbf{h} \in \mathbb{C}^{QS\times 1}$) for a URLLC device at a random position within the area. To make a comparison, a centralized mMIMO deployment is also considered, where there are $M = QS = 400$ antennas deployed at the centre of the area. As shown in the figure, compared to mMIMO, most devices in the area can achieve much larger channel gains in CF mMIMO thanks to the macro-diversity. In particular, considering the maximum channel gain that is achieved with a probability of $1 - 10^{-5}$, i.e., the 99.999% channel gain availability, a more than 25 dB enhancement can be provided by CF mMIMO, which certainly increases the received power of the desired signal, leading to a lower decoding error probability as a result.

2) Signal Spatial Sparsity: In addition, since the signal of a device to different APs undergoes different levels of large-scale fading, the neighboring APs in the vicinity of a device usually capture more significant signal energy than other APs. As a result, only neighboring APs within a communication range of a device have non-negligible channel gains thanks to macro-diversity [110], which leads to the signal spatial sparsity in CF mMIMO [112]. The salient feature is that if an appropriate set of APs is chosen for each device and the sets of APs for different devices are different or partially overlapped, and the interference can be reduced significantly, unlike mMIMO, resulting in significantly reduced probability of preamble collision.

TABLE II
RATIO OF CHANNEL GAIN OBTAINED BY $Q_{\rm ap}$ NEIGHBOURING APS OF A DEVICE AND TOTAL CHANNEL GAIN BY ALL APS.

$Q_{\rm ap}$	1	4	8	16
Ratio	40.0%	70.1%	80.7%	89.5%

To illustrate this feature, we present the ratio of channel gain obtained by the $Q_{\rm ap}$ APs closest to a device and that by all APs in Table II. The simulation scenario and setup are the same as those in Fig. 13. As we can see in the results, on average, 40.0% signal energy of a device can be captured by a single AP that is closest to a device, and over 80% can be captured by only $Q_{\rm ap} = 8$ closest APs. Thus, employing a certain number of neighboring APs of a device to serve it in CF mMIMO may only cause a small channel gain loss. Furthermore, we note that it requires less fronthaul overhead for signal processing than employing all APs and can help in reducing the multi-user interference from devices that are far away from it [108].

B. Research Directions and Challenges

The distinctive features of CF mMIMO open up new avenues for resolving preamble collision and suppressing multi-user interference in GF random access, which thus help with the support of URLLC. In the following, we shed light on the potential research directions as well as challenges for GF URLLC in CF mMIMO.

1) Preamble-Collision Resolution: One of the important research directions is preamble-collision resolution design in GF URLLC with CF mMIMO. In mMIMO, as explained in Section IV-B3, preamble-collision resolution for GF URLLC could be achieved by coded random access and its variants in the context of HARQ retransmission schemes, which rely on the use of multiple preamble transmissions over multiple transmission TTIs. Going one step further, by exploiting macro-diversity and signal spatial sparsity, CF mMIMO is capable of resolving preamble collision on the basis of a single transmission TTI, which is beyond the capability of mMIMO [113].

To provide an insight into this, we use a toy example of a two-device preamble collision scenario in GF random access, in which URLLC device 1 and URLLC device 2 are active and select the same preamble in a transmission TTI. Furthermore, they are assumed to be far away geographically, as illustrated in Fig. 12 (this is sensible since devices become active independently and select preambles randomly). In such a scenario, their estimated noise-free CSI over all the APs can be written as

$$\hat{\mathbf{h}} = \mathbf{h}_1 + \mathbf{h}_2. \tag{7}$$

As the estimated CSI is a superposition of the CSI of two devices, neither of their data could be decoded based on it. Note that in centralized mMIMO, since all the signals of collided devices are multiplexed at the centralized BS, the BS can only deem that $\hat{\mathbf{h}}$ is a single device's CSI, making it unable to resolve the preamble collision regardless of the locations of collided devices. In CF mMIMO, on the other hand, since the two devices are far away from each other, they can be surrounded by different sets of neighboring APs, denoted by $\mathcal{Q}_1$ and $\mathcal{Q}_2$, respectively, where the cardinality $|\mathcal{Q}_1| = |\mathcal{Q}_2| = Q_{\rm ap}$. Thanks to macro-diversity and signal spatial sparsity, the strength of received signals from one device is negligible at the neighboring APs of the other, which means that $\mathbf{h}_{1,\mathcal{Q}_2} \approx \mathbf{0}$ and $\mathbf{h}_{2,\mathcal{Q}_1} \approx \mathbf{0}$, where $\mathbf{h}_{i,\mathcal{Q}_j}$ represents the channel vector of device i over APs in $\mathcal{Q}_j$. In the light of this, if the BS only employs APs in $\mathcal{Q}_1$ to serve device 1, then its estimated CSI can be approximately written as

$$\hat{\mathbf{h}}_{1,\mathcal{Q}_1} = \mathbf{h}_{1,\mathcal{Q}_1} + \mathbf{h}_{2,\mathcal{Q}_1} \approx \mathbf{h}_{1,\mathcal{Q}_1}. \tag{8}$$

Similarly, we have $\hat{\mathbf{h}}_{2,\mathcal{Q}_2} \approx \mathbf{h}_{2,\mathcal{Q}_2}$, i.e., $\mathbf{h}_{1,\mathcal{Q}_2}$ can be ignored. These indicate that the estimated CSI of device i over APs in $\mathcal{Q}_i$ approximately becomes interference-free from preamble collision, which indicates the potential of CF mMIMO in preamble collision resolution in GF random access. To verify this advantage, we compare the SINR of device 1 in mMIMO and CF mMIMO in the two-device preamble collision scenario as shown in Fig. 14. The simulation setup is the same as those in Fig. 13. The two devices are at random positions within the area, and $Q_{\rm ap}$ is set to 4. Evidently, in the context of preamble collision, the SINR of the collided device in CF mMIMO with $Q_{\rm ap} = 4$ can be significantly higher than that in mMIMO and in CF mMIMO with all APs, which thus leads to a lower decoding error probability.

Obviously, the benefit of preamble-collision resolution is greatly helpful in enhancing the contention-based URLLC transmissions in GF random access. For instance, in the reactive HARQ retransmission scheme, this benefit facilitates

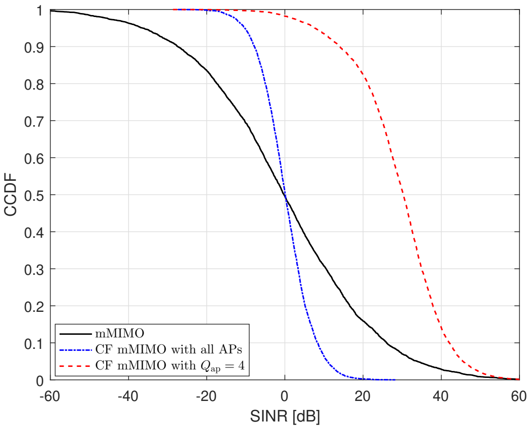


Fig. 14. SINR comparison of collided device in mMIMO and CF mMIMO when two devices select the same preamble in GF random access.

more devices with preamble collision to be served data in each single transmission TTI. This not only reduces their access latency and need for retransmissions, but also in turn, leads to contention alleviation in the upcoming transmission/retransmission TTIs. Nevertheless, designing feasible schemes in GF URLLC that can fully exploit this outstanding phenomenon is worth investigating. In particular, in the URLLC scenario with sporadic and unpredictable traffic patterns, the BS may have neither prior information of devices' activity and locations nor their preamble choices. In such a case, how to acquire the neighboring APs for devices with preamble collision and determine the optimal Q_{ap} and other necessary conditions to satisfy reliability and latency requirements are challenging. In [114], a framework solution based machine learning-enabled K-means AP clustering algorithm was proposed to provide some insights into addressing the challenge. Nevertheless, the impact of this framework solution on the key metrics of URLLC, such as reliability and latency as well as access load, is not studied. In addition, in K-repetition or proactive HARQ retransmission, how to effectively employ preamble-collision resolution over multiple consecutive transmissions to maximize the access load under given reliability and latency constraints needs to be explored.

2) Channel Inference Assisted Resource Allocation: In some URLLC use cases, the traffic pattern of devices could be deterministic and predictable [4]. In such a case, the channel statistics or even locations of active devices can be prior known at the BS. Using this information, employing optimal resource allocation, such as devices' preambles, transmit power, and associated AP subsets plays a vital role in contention-based GF URLLC to minimize/control the preamble-collision effect well as multi-user interference [4] [115]. Nevertheless, due to the limited preamble resources, the optimal resource allocation does not necessarily ensure to suppress the preamble collision as well as multi-user interference to the desired level so as to meet the stringent reliability constraint within the target latency. To enhance it, a potential research direction is to exploit the benefits of channel inference to make a more efficient resource utilization in GF URLLC [39].

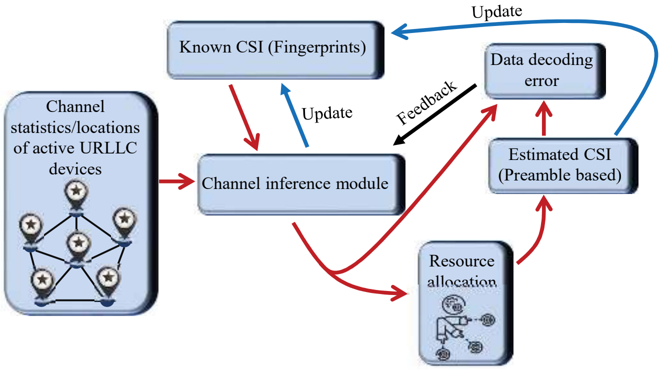


Fig. 15. An illustration of channel inference assisted resource allocation process. The inference module is able to predict the new CSI based on known CSI and to be reinforced based on the feedback of resulting data decoding error.

The key idea of channel inference is to infer the CSI of active devices in the current random-access slot based on their nearby devices' CSI obtained in previous slots (within the coherent time) [116]. Advanced artificial intelligence techniques can be employed to enable inference functionality [39]. To be specific, among the active URLLC devices in a random access slot, some of them may have close locations to the active devices in the previous slots. Since the channel correlation between nearby devices is usually high in terms of large-scale fading, if the CSI of devices in the previous slots are obtained at the BS, the BS then can utilize channel inference module, as illustrated in Fig. 15, to predict the CSI of the current devices based on the information of channel statistics, relative locations between devices, and environment parameters etc. For the devices with predicted CSI, they can directly transmit data packets without sending preambles. As a result, one significant resulting benefit is that there are fewer devices competing for the limited preamble resource in GF random access. Therefore, more efficient resource allocation and utilization can be achieved. Nevertheless, a generalized framework for resource allocation subject to scalability-reliability-latency tradeoff in CF mMIMO is still missing and devising feasible algorithms in CF mMIMO for such inference-based resource allocation further complicates the problem [117].

3) Limited Fronthaul: As shown in Fig. 12, fronthaul links are used to connect distributed APs to the CPU. The optical fibre may be an ideal candidate for fronthaul links due to its high channel capacity. However, for a large number of APs, the deployment cost becomes prohibitively high. In this case, low-cost alternatives , e.g., hybrid RF/free space optic (FSO) [118], can be considered. The resulting fronthaul links have limited capacity compared to those with optical fibre. Thus, bandwidth-efficient approaches to compress and forward the CSI and data of devices from APs to the CPU are studied [119]–[121].

In URLLC, in fact, any delay from APs to the CPU due to processing at APs, including scheduling is undesirable and should be minimized. To this end, in the resource allocation for fronthaul links, the processing delay has to be taken into account to meet the latency constraints, especially for GF random access. However, since the number of devices is large and the activity of devices with sporadic traffic is difficult to predict, scheduling and resource allocation to access fronthaul links become challenging.

VI. Conclusions

In this survey paper, we first outlined typical URLLC use cases and requirements and presented an overview of essential enabling techniques and principles of GF URLLC in NR. We identified and discussed key factors such as preamble collision and multi-user interference that jeopardize transmission reliability within a target latency in GF URLLC. Once we identified those factors, we then revisited the phenomena of mMIMO, i.e., channel hardening and favorable propagation, and discussed how to enhance the performance of GF URLLC by exploiting mMIMO. Finally, we shed light on the distinctive features and potential benefits of CF mMIMO over centralized mMIMO and provided a perspective on its potential research directions along with challenges towards enabling GF URLLC.

References

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[2] J. Ding, M. Nemati, C. Ranaweera, and J. Choi, "IoT connectivity technologies and applications: A survey," *IEEE Access*, vol. 8, pp. 67646–67673, 2020.

[3] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796–131813, 2020.

[4] P. Popovski, Č. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[6] K. S. Kim, D. K. Kim, C. Chae, S. Choi, Y. Ko, J. Kim, Y. Lim, M. Yang, S. Kim, B. Lim, K. Lee, and K. L. Ryu, "Ultrareliable and low-latency communication techniques for tactile Internet services," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 376–393, 2019.

[7] NGMN Alliance, "5G E2E technology to support verticals URLLC requirements," *NGMN Alliance*, 2020.

[8] 3GPP, "Service requirements for the 5G system (release 16)," *3GPP TS 22.261*, Sep. 2018.

[9] 3GPP, "Study on latency reduction techniques for LTE (release 14)," *3GPP TR 36.881*, May 2016.

[10] 3GPP, "Study on scenarios and requirements for next generation access technologies (release 14)," *3GPP TSG RAN, TR 38.913*, May 2017.

[11] 5G Americas, "The 5G evolution: 3GPP releases 16-17," *5G Americas*, 2020. https://www.5gamericas.org/5g-evolution-3gpp-releases-16-17.

[12] Lenovo and Motorola Mobility, "HARQ design for uplink grant-free transmission," *3GPP, R1-1717857*, Oct. 2017.

[13] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); physical layer procedures," *3GPP Standard TS 36.213, V14.2.0*, Mar. 2017.

[14] G. Berardinelli, N. Huda Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23602–23611, 2018.

[15] 3GPP RP-181477, "SID on physical layer enhancements for NR URLLC," Jun. 2018.

[16] N. H. Mahmood and H. Alves, "Dynamic multi-connectivity activation for ultra-reliable and low-latency communication," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 112–116, August 2019.

[17] J. Choi, "On fast retrial for two-step random access in MTC," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1428–1436, 2021.

[18] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. Poor, *MIMO Wireless Communications*. Cambridge University Press, 2007.

[19] K. K. Wong, R. D. Murch, and K. B. Letaief, "Performance enhancement of multiuser MIMO wireless communication systems," *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 1960–1970, 2002.

[20] C. Oestges and B. Clerckx, *MIMO Wireless Communications: From Real-World Propagation to Space-Time Code Design*. Elsevier Science, 2010.

[21] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, pp. 3590–3600, November 2010.

[22] T. Marzetta, *Fundamentals of Massive MIMO*. Fundamentals of Massive MIMO, Cambridge University Press, 2016.

[23] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, pp. 186–195, February 2014.

[24] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Transactions on Communications*, vol. 61, pp. 1436–1449, April 2013.

[25] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 1293–1308, Feb 2016.

[26] S. Jin, J. Wang, Q. Sun, M. Matthaiou, and X. Gao, "Cell coverage optimization for the multicell massive MIMO uplink," *IEEE Transactions on Vehicular Technology*, vol. 64, pp. 5713–5727, Dec 2015.

[27] H. Han, X. Guo, and Y. Li, "A high throughput pilot allocation for M2M Communication in Crowded Massive MIMO Systems," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 9572–9576, Oct 2017.

[28] F. Ahsan and A. Sabharwal, "Leveraging massive MIMO spatial degrees of freedom to reduce random access delay," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 2007–2011, Oct 2017.

[29] S. R. Panigrahi, N. Bjorsell, and M. Bengtsson, "Feasibility of large antenna arrays towards low latency ultra reliable communication," in *2017 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1289–1294, March 2017.

[30] W. Tarneberg, M. Karaca, A. Robertsson, F. Tufvesson, and M. Kihl, "Utilizing massive MIMO for the tactile Internet: Advantages and trade-offs," in *2017 IEEE International Conference on Sensing, Communication and Networking (SECON Workshops)*, pp. 1–6, June 2017.

[31] A. Bana, E. de Carvalho, B. Soret, T. Abrão, J. C. Marinello, E. G. Larsson, and P. Popovski, "Massive MIMO for Internet of Things (IoT) connectivity," *Physical Communication*, vol. 37, p. 100859, 2019.

[32] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[33] G. Interdonato, G. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free Massive MIMO communications," *EURASIP Journal on Wireless Communications and Networking*, vol. 197, pp. 1–13, 2019.

[34] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878–99888, 2019.

[35] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133995–134030, 2020.

[36] S. Venkatesan, A. Lozano, and R. Valenzuela, "Network MIMO: Overcoming intercell interference in indoor wireless systems," in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, pp. 83–87, November 2007.

[37] T. K. Le, U. Salim, and F. Kaltenberger, "An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17," *IEEE Access*, vol. 9, pp. 433–444, 2021.

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

[39] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 204–246, 2021.

[40] 3GPP, "Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC) (release 16)," *3GPP TR 38.824 V2.0.1*, 2019.

[41] 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, 2017.

[42] S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4734–4746, 2020.

[43] A. Vergutz, G. Noubir, and M. Nogueira, "Reliability for smart healthcare: A network slicing perspective," *IEEE Network*, vol. 34, no. 4, pp. 91–97, 2020.

[44] D. Feng, C. She, K. Ying, L. Lai, Z. Hou, T. Q. S. Quek, Y. Li, and B. Vucetic, "Toward ultra-reliable low-latency communications: Typical scenarios, possible solutions, and open issues," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 94–102, 2019.

[45] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 1–6, August 2018.

[46] A. Mahmood, M. I. Ashraf, M. Gidlund, and J. Torsner, "Over-the-air time synchronization for URLLC: Requirements, challenges and possible enablers," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 1–6, August 2018.

[47] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, 2018.

[48] T. Fehrenbach, R. Datta, B. Göktepe, T. Wirth, and C. Hellge, "URLLC services in 5G low latency enhancements for LTE," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pp. 1–6, August 2018.

[49] 5G Americas, "New services and applications with 5G ultra-reliable low-latency communication," *5G Americas*, 2020. https://www.5gamericas.org/new-services-applications-with-5g-ultra-reliable-low-latency-communications/.

[50] S. R. Pokhrel and S. Garg, "Multipath communication with deep Q-Network for industry 4.0 automation and orchestration," *IEEE Transactions on Industrial Informatics*, 2020.

[51] T. N. Weerasinghe, I. A. Balapuwaduge, and F. Y. Li, "Priority-based initial access for URLLC traffic in massive IoT networks: Schemes and performance analysis," *Computer Networks*, vol. 178, p. 107360, 2020.

[52] J. Thota and A. Aijaz, "On performance evaluation of random access enhancements for 5G uRLLC," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–7, 2019.

[53] S. R. Pokhrel and J. Choi, "Improving TCP performance over WiFi for Internet of vehicles: A federated learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6798–6802, 2020.

[54] X. Yang, L. Liu, N. H. Vaidya, and F. Zhao, "A vehicle-to-vehicle communication protocol for cooperative collision warning," in *The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004.*, pp. 114–123, IEEE, 2004.

[55] M. Gupta, J. Benson, F. Patwa, and R. Sandhu, "Secure V2V and V2I communication in intelligent transportation using cloudlets," *IEEE Transactions on Services Computing*, pp. 1–1, 2020.

[56] M. Nemati and H. Arslan, "Low ICI Symbol Boundary Alignment for 5G Numerology Design," *IEEE Access*, vol. 6, pp. 2356–2366, 2018.

[57] M. I. Ashraf, Chen-Feng Liu, M. Bennis, and W. Saad, "Towards low-latency and ultra-reliable vehicle-to-vehicle communication," in *2017 European Conference on Networks and Communications (EuCNC)*, pp. 1–5, June 2017.

[58] S. A. Busari, M. A. Khan, K. M. S. Huq, S. Mumtaz, and J. Rodriguez, "Millimetre-wave massive MIMO for cellular vehicle-to-infrastructure communication," *IET Intelligent Transport Systems*, vol. 13, no. 6, pp. 983–990, 2019.

[59] G. Naik, J. Liu, and J. J. Park, "Coexistence of dedicated short range communications (DSRC) and Wi-Fi: Implications to Wi-Fi performance," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*.

[60] S. R. Pokhrel, "Federated learning meets blockchain at 6G edge: A drone-assisted networking for disaster response," in *Proceedings of the 2nd ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond*, pp. 49–54, September 2020.

[61] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile internet of things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2995–3005, 2019.

[62] A. Kumbhar, F. Koohifar, I. Güvenç, and B. Mueller, "A survey on legacy and emerging technologies for public safety communications," *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 97–124, 2017.

[63] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[64] 3GPP, "General packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access (release 15)," *3GPP TS23.401 V15.4.0*, 2018.

[65] ZTE Microelectronics, "On 2-step random access procedure," *R1-1608969, 3GPP TSG RAN WG1 Meeting 86b*, 2016.

[66] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of lte and lte-a suitable for m2m communications? a survey of alternatives," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 4–16, First 2014.

[67] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Communications*, vol. 24, pp. 120–128, February 2017.

[68] IEEE Standard for Local and metropolitan area networks, *Part 16: Air Interface for Broadband Wireless Access Systems*, May 2009.

[69] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 2861–2864, September 2007.

[70] 3GPP, "Semi-persistent scheduling for 5G new radio URLLC," *R1-167309, 3GPP TSG-RAN WG1*, Aug. 2016.

[71] E. Fitzgerald and M. Pióro, "Efficient pilot allocation for URLLC Traffic in 5G Industrial IoT Networks," in *2019 11th International Workshop on Resilient Networks Design and Modeling (RNDM)*, pp. 1–7, October 2019.

[72] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Preamble reservation based access for grouped mMTC Devices with URLLC requirements," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2019.

[73] 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)*, pp. 1–6, December 2017.

[74] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1005–1010, May 2017.

[75] 3GPP TSG RAN WG1, "RAN1 Chairman's Notes," Jan. 2017.

[76] 3GPP TR-RAN1, "Discussion on HARQ support for URLLC," *R1-1612246*, 2016.

[77] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 607–612, August 2019.

[78] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 741–755, 2021.

[79] J. H. Sørensen, E. de Carvalho, C. Stefanovic, and P. Popovski, "Coded pilot random access for massive MIMO systems," *IEEE Trans. Wireless Communications*, vol. 17, no. 12, pp. 8035–8046, 2018.

[80] D. Bertsekas and R. Gallager, *Data Networks (2nd Ed.)*. USA: Prentice-Hall, Inc., 1992.

[81] Y. Choi, Suho Park, and Saewoong Bahk, "Multichannel random access in OFDMA wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 603–613, 2006.

[82] J. Choi, "On throughput improvement using immediate re-transmission in grant-free random access with massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8341–8350, 2020.

[83] 3GPP, "3GPP TS 38.300 V15.5.0: NR and NG-RAN overall description; Stage 2; (Release 15)," *Tech. Rep.*, Mar. 2019.

[84] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, 2014.

[85] P. Yang, Y. Xiao, M. Xiao, and S. Li, “6G wireless communications: Vision and potential techniques,” *IEEE Network*, vol. 33, no. 4, pp. 70–75, 2019.

[86] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, “Multiple-antenna channel hardening and its implications for rate feedback and scheduling,” *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1893–1909, 2004.

[87] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Aspects of favorable propagation in massive MIMO,” in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 76–80, 2014.

[88] F. A. P. De Figueiredo, F. A. C. M. Cardoso, I. Moerman, and G. Fraidenraich, “On the application of massive MIMO systems to machine type communications,” *IEEE Access*, vol. 7, pp. 2589–2611, 2019.

[89] J. Hoydis, S. ten Brink, and M. Debbah, “Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, 2013.

[90] 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, 2010.

[91] 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, 2016.

[92] S. Schiessl, J. Gross, M. Skoglund, and G. Caire, “Delay Performance of the Multiuser MISO Downlink Under Imperfect CSI and Finite-Length Coding,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 765–779, 2019.

[93] J. Choi, *Optimal combining and detection: statistical signal processing for communications*. Cambridge University Press, 2010.

[94] H. Yan, A. Ashikhmin, and H. Yang, “Can massive MIMO support URLLC?,” in *2021 93rd Vehicular Technology Conference*, pp. 1–5, Oct Apr. 2021.

[95] S. Shahsavari, A. Ashikhmin, E. Erkip, and T. L. Marzetta, “Coordinated multi-point massive MIMO cellular systems with sectorized antennas,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 2130–2135, Oct. 2018.

[96] J. Ding, D. Qu, H. Jiang, and T. Jiang, “Success probability of grant-free random access with massive MIMO,” *IEEE Internet of Things Journal*, vol. 6, pp. 506–516, Feb 2019.

[97] S. Sesia, M. Baker, and I. Toufik, *LTE - The UMTS Long Term Evolution: from theory to practice*. John Wiley & Sons, 2011.

[98] J. Ding, D. Qu, M. Feng, J. Choi, and T. Jiang, “Dynamic preamble-resource partitioning for critical MTC in massive MIMO Systems,” *IEEE Internet of Things Journal*, pp. 1–12, 2021.

[99] Q. Zhang, S. Jin, and H. Zhu, “A hybrid-grant random access scheme in massive MIMO Systems for IoT,” *IEEE Access*, vol. 8, pp. 88487–88497, 2020.

[100] H. Han, Y. Li, W. Zhai, and L. Qian, “A grant-free random access scheme for M2M Communication in Massive MIMO Systems,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3602–3613, 2020.

[101] J. Ding and J. Choi, “Triangular non-orthogonal random access in mMIMO systems,” *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6918–6931, 2020.

[102] Y. Chen, L. Cheng, and L. Wang, “Prioritized resource reservation for reducing random access delay in 5G URLLC,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, Oct. 2017.

[103] J. Choi, “An approach to preamble collision reduction in grant-free random access with massive MIMO,” *IEEE Trans. Wireless Communications*, vol. 20, no. 3, pp. 1557–1566, 2021.

[104] J. Ding and J. Choi, “On outage probability for grant-free random access with K-repetition transmission,” *(submitted)*, pp. 1–5, 2021.

[105] H. Jiang, D. Qu, J. Ding, and T. Jiang, “Multiple preambles for high success rate of grant-free random access with massive MIMO,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4779–4789, 2019.

[106] G. Casciano, P. Baracca, and S. Buzzi, “Enabling ultra reliable wireless communications for factory automation with distributed MIMO,” in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, pp. 1–7, Sep. 2019.

[107] E. Björnson and L. Sanguinetti, “Making cell-free massive MIMO competitive with MMSE processing and centralized implementation,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2020.

[108] S. Buzzi and C. D’Andrea, “Cell-free massive MIMO: User-centric approach,” *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 706–709, 2017.

[109] P. Liu, K. Luo, D. Chen, and T. Jiang, “Spectral efficiency analysis of cell-free massive MIMO systems with zero-forcing detector,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 795–807, 2020.

[110] Z. Chen and E. Björnson, “Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry,” *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5205–5219, 2018.

[111] 3GPP, “Scenarios, frequencies and new field measurement results from two operational factory halls at 3.5 GHz for various antenna configurations,” *3GPP R1-1813177*, Nov. 2018.

[112] J. Choi, “Compressive random access for MTC in distributed input distributed output systems,” in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, June 2017.

[113] J. Choi, “Multi-channel estimation of devices with preamble collision in distributed MTC.” Manuscript submitted for publication to IEEE SSP’21, 2021.

[114] J. Ding, D. Qu, P. Liu, and J. Choi, “Machine learning enabled preamble collision resolution in distributed massive MIMO,” *IEEE Transactions on Communications*, pp. 1–1, 2021.

[115] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, “Structured massive access for scalable cell-free massive MIMO systems,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1086–1100, 2021.

[116] H. Khan, M. M. Butt, S. Samarakoon, P. Sehier, and M. Bennis, “Deep learning assisted CSI estimation for joint URLLC and eMBB resource allocation,” in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, June 2020.

[117] P. N. Alevizos, E. A. Vlachos, and A. Bletsas, “Inference-based distributed channel allocation in wireless sensor networks,” *arXiv preprint arXiv:1703.06652*, 2017.

[118] M. Najafi, V. Jamali, D. W. K. Ng, and R. Schober, “C-RAN with hybrid RF/FSO fronthaul links: Joint optimization of fronthaul compression and RF time allocation,” *IEEE Trans. Communications*, vol. 67, no. 12, pp. 8678–8695, 2019.

[119] W. Xia, J. Zhang, T. Q. S. Quek, S. Jin, and H. Zhu, “Joint optimization of fronthaul compression and bandwidth allocation in uplink H-CRAN with large system analysis,” *IEEE Trans. Communications*, vol. 66, no. 12, pp. 6556–6569, 2018.

[120] F. Riera-Palou and G. Femenias, “Decentralization issues in cell-free massive MIMO networks with zero-forcing precoding,” in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 521–527, Sep. 2019.

[121] H. Masoumi and M. J. Emadi, “Performance analysis of cell-free massive MIMO system with limited fronthaul capacity and hardware impairments,” *IEEE Trans. Wireless Communications*, vol. 19, no. 2, pp. 1038–1053, 2020.