SCIENCE CHINA
Information Sciences

• REVIEW •

September 2022, Vol. 65 191301:1–191301:45
https://doi.org/10.1007/s11432-021-3470-1

A new 5G radio evolution towards 5G-Advanced

Jiyong PANG[1*], Shaobo WANG[1], Zhenfei TANG[1], Yanmin QIN[1], Xiaofeng TAO[2], Xiaohu YOU[3,4] & Jinkang ZHU[5]

[1]*Research Department, Wireless Network, Huawei Technologies, Shanghai* 201206*, China;*
[2]*National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing* 100876*, China;*
[3]*National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing* 210096*, China ;*
[4]*Purple Mountain Laboratories, Nanjing* 211111*, China;*
[5]*School of Information Science and Technology, University of Science and Technology of China, Hefei* 230026*, China*

Received 15 November 2021/Revised 8 February 2022/Accepted 2 April 2022/Published online 29 August 2022

Abstract The evolution of the fifth-generation (5G) new radio (NR) has progressed swiftly since the third generation partnership project (3GPP) standardized the first NR version (Release 15) in mid-2018. Nowadays, the world's leading carriers are competing to provide various commercial services over 5G networks. Looking ahead to 2025 and beyond, it is expected that over 6.5 million 5G base stations will be installed to offer services to over 58% of the world's population via over 100 billion 5G connections. Following the rapid development of 5G, an increasing number of commercialization use cases will drive the 5G network to continuously improve performance and expand capabilities. Hence, it is the right time to consider a well-defined framework and standardization for 5G NR evolution (5G-Advanced) to support commercialization between 2025 and 2030. First, this study addresses the key driving forces, requirements, usage scenarios, and capabilities of 5G-Advanced; then, it highlights the main technological challenges and introduces the top 10 promising technological directions in detail. Finally, other fascinating technological directions in 5G-Advanced are shortly mentioned.

Keywords 3GPP, 5G, NR, 5G-Advanced

Citation Pang J Y, Wang S B, Tang Z F, et al. A new 5G radio evolution towards 5G-Advanced. Sci China Inf Sci, 2022, 65(9): 191301, https://doi.org/10.1007/s11432-021-3470-1

1 Introduction

The standardization of the 5G mobile communication system started in 2016 under the framework of the third generation partnership project (3GPP), which includes 5G core network and 5G new radio (5G NR). As the goal for IMT-2020, the international telecommunication union (ITU) identified three major communication scenarios for 5G; they include enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low latency communications (URLLC), which represent three types of services with distinct performance requirements [1]. The first full release of 3GPP's 5G standards was frozen with Release 15 (Rel-15) in December 2017 and in June 2018 for the non-standalone and standalone variants, respectively, with a focus on eMBB. The subsequent Release 16 (Rel-16) and Release 17 (Rel-17) focus primarily on URLLC and mMTC to expand the availability and applicability of 5G in both the industrial Internet of Things (IIoT) and vertical industries. However, Rel-16 was frozen in March 2020 and Rel-17 is expected to be completed in March 2022 [2–4].

Meanwhile, 5G networks are globally deployed and commercially launched at an unprecedented speed. Compared with 2019, the number of commercial 5G networks, terminal types, and users has increased by 20, 21, and 350 times, respectively, at the end of April 2021, registering over 160 networks, 700 terminal types, and 350 million users [5]. User experience and advanced industrial digitalization have both been revolutionized by 5G. Extended reality (XR)[1)] users have increased by 35% annually, and over 5000 5G

* Corresponding author (email: pangjiyong@huawei.com)

1) XR is an umbrella term that groups augmented, virtual, and mixed reality into one term.

industrial digitalization projects such as 5G port and steel have formed the market size effect [6, 7]. It is expected that over 6.5 million 5G base stations (BSs) will be deployed to provide 5G services to 58% of the world's population, with 337 million people using the XR, and the number of global 5G connections will exceed 100 billion by 2025 [8, 9]. The rapid commercial deployment of 5G networks and the maturity of the 5G devices will spawn several new applications and triggers an explosion of mobile data traffic in both ToC (to consumer) and ToB (to business) areas.

As 5G continues to attain new commercialization milestones and the 5G industry produces new technologies, new frontiers of 5G are revealed. However, the current 5G network will be insufficient to meet new requirements such as the personal consumer experience upgrades and digital transformation of the industry in the next decade. Thus, this is an appropriate time to consider the continuous evolution and enhancement of 5G toward 2025 and beyond to empower the future. In November 2020, Huawei sketched out plans for 5.5G [10], and then at the 46th project co-ordination group meeting of the 3GPP held in April 2021, 5G evolution for providing stronger capabilities was officially named 5G-Advanced and Release 18 (Rel-18) was recognized as the first release of 5G-Advanced [11]. After a six-month intense discussion, Rel-18 work packages were approved in December 2021 at the 3GPP radio access network plenary meeting [4]. Before the sixth generation (6G) arrives at the end of the decade, 5G-Advanced will be a key focus for both 3GPP and industries [12–17].

In this study, we explain our knowledge of 5G-Advanced from an industry-standard and academic research perspective. The remainder of this study is organized as follows: Section 2 describes the driving forces, usage scenarios, key performance indicators (KPIs), and promising technological directions of 5G-Advanced. Sections 3, 4, and 5 highlight in detail the potential technologies in each direction. Finally, Section 6 concludes the study.

2 5G-Advanced vision

The world has witnessed new mobile generations in every decade since the 1980s; there is occasionally a mid-decade upgrade during each generation, which has already happened to the third and fourth generations (3G and 4G). Similarly, as a key update of the 5G specifications, 5G-Advanced will begin with Rel-18 in the middle of this decade, i.e., around 2025, and will further improve in Rel-19 and 20 onwards to fully deliver the 5G expectation. Simultaneously, 5G-Advanced is expected to introduce several technologies that will eventually be included in 6G, and this is considered an intermediate step between 5G and future 6G. In this section, we will provide our view of 5G-Advanced starting with the driving forces and the valuable application scenarios. Furthermore, we will introduce KPIs and promising technological directions.

2.1 Driving forces

The 5G industry has identified five factors that enable the 5G-Advanced vision. First, lessons learned from 5G commercial deployment must be adequately addressed. Second, existing 5G capabilities must be strengthened to meet future ToC and ToB requirements. Third, for immersive applications such as XR and holograms, high-speed real-time experience is required. Fourth, uplink (UL)-centric networks must be built to provide ten times the UL speed for industrial digitalization. Finally, 5G network capabilities must be expanded to support new business opportunities.

2.1.1 *Addressing commercial deployment challenges*

During the rapid deployment of 5G networks, several lessons were learned that significantly affected the commercial success of 5G. From the capital expenditure (CapEx) viewpoint, spectrum auction and site acquisition will cost more than 4G mainly because of the wider bandwidth required and limited infrastructure capacity, respectively. Additionally, 5G equipment is more expensive because of more advanced functionality and radio frequency components. From the operating expense (OpEx) viewpoint, higher power consumption and more complicated operation and maintenance are the two major expenses. Thus, another objective of 5G-Advanced is to address existing challenges in 5G commercial deployment and operation. Continuous technological evolution will be an essential tool for reducing the CapEx by improving the performance-price ratio and OpEx by adopting better and more intelligent solutions.

2.1.2 *Existing 5G capabilities enhancement*

Personal data consumption is increasing rapidly for eMBB. The data usage of an individual user per month is expected to reach 150 GB by 2025, which is a significant increase from 10 and 4 GB in 2020 and 2018, respectively, with a compound annual growth rate of 65% [18]. For mMTC, narrowband IoT (NB-IoT) has already been integrated into NR to support low-speed, low-power services. Furthermore, 3GPP Rel-17 is currently defining a new reduced-capability (RedCap) terminal type called RedCap new radio to support medium-rate IoT services. However, because of the gap in supporting some higher-rate IoT services, there might be a need to define even more terminal types to meet the requirements of diversified MTC services. For URLLC, 5G enhancement capabilities do not only satisfy its low latency requirement but also guarantee a certain latency to support wireless deterministic and time-sensitive communication applications such as remote operation and programmable logic controller [19].

2.1.3 *Adoption of XR-pro immersive experience*

As new applications emerge in the market, 5G capability boundaries must be expanded to open up a new space of business. With the maturity of the terminal side, immersive interaction services represented by XR are rapidly developing and will become mainstream services in the next two or three years. Additionally, providing services such as high-definition (HD) interactive holograms at any time and place will be widely adopted in the future. To deliver seamless connectivity for such applications, 5G network must be capable of providing high-bandwidth access with guaranteed low latency. For example, 5G network must increase its average speed from 120 megabytes per second (Mbps) to 2 gigabytes per second (Gbps) as the standard of HD improves from 4 to 16 K resolution. To guarantee real-time interaction in virtual worlds, 5G must further reduce the transmission latency from the current 20 to 5 ms [10, 20].

2.1.4 *UL centric services*

With the acceleration of the entire society's digital transformation, the demand for UL business in ToC, ToB, and IoT fields has expanded, creating significant challenges to 5G networks for both UL capacity and coverage. In the ToC field, diverse consumer XR services require HD images and videos to be uploaded from local devices to clouds for further rendering. Multimedia social networks have led to a proliferation of user-generated content sent to social media platforms, particularly at transport hubs, malls, tourist attractions, sports stadiums, and other locations of special interest. In the ToB field, digital exploration in steel, mining, port, manufacturing, and education shows that video surveillance, remote control, and machine vision are typical industrial applications, where uploading HD and ultra-high definition (UHD) photos and videos place high requirements on the network's UL capacity. In the IoT field, massive broadband IoT devices such as wireless cameras are deployed for wide/outdoor area surveillance, vehicle monitoring, and unmanned vehicle distribution, which exert pressure on the overall network UL capacity. All these lead to a significant increase of over 40% in the proportion of UL traffic over 5G networks. To cope with such huge UL traffic, 5G UL capabilities must be improved by at least ten times [10, 20].

2.1.5 *Diversified industrial IoT scenarios*

Broadly enabling industrial applications is a significant feature that distinguishes 5G from the previous mobile communication technologies. Presently, the accelerated adoption of 5G worldwide further boosts the integration of 5G with a wide range of vertical applications such as smart manufacturing, ports, mines, and smart energy. Among these applications, several location-based services play an important role during the enterprise digital transformation, where extremely high positioning accuracy of 10 cm or less and latency in the order of 10 ms is highly desirable. Additionally, 5G must provide sensing capabilities with an adaptive resolution to enable remote control, vehicle-road synergy, unmanned transportation, and smart logistics. For example, the introduction of beam sweeping through cellular massive multiple-input and multiple-output (MIMO) into sensing technologies such as sensing, communication, and even indoor high-precision positioning services, can be implemented over wireless technologies. In summary, 5G connectivity should support diversified IoT capabilities such as narrowband/broadband connectivity, high-precision positioning, and sensing IoT.

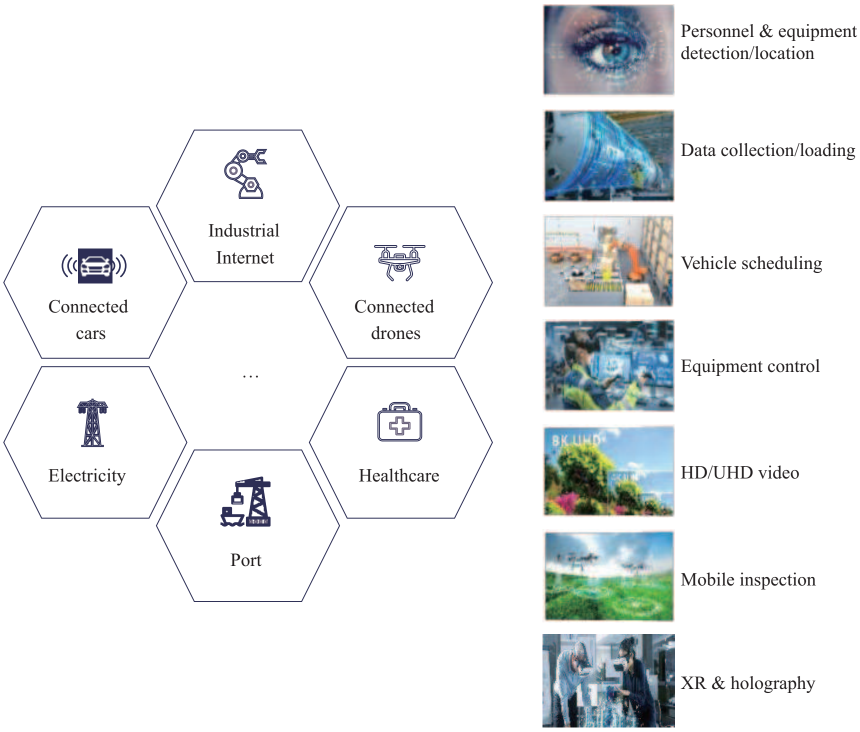

Figure 1 (Color online) Diversified IIoT scenarios and their service flows.

2.2 Usage scenarios

The current 5G scenario is an upgrade and extension of 5G-Advanced. In response to the above-mentioned driving forces, the eMBB, mMTC, and URLLC standard scenarios set by ITU for 5G will be enhanced and extended in 5G-Advanced. Further, eMBB and fixed wireless access will help enable smart homes, video everywhere, and XR experiences. RedCap will be added to mMTC to support the different types of devices essential for broadband IoT. Latency will be more restricted in URLLC to support connectivity requirements for smart manufacturing (remote movement control requires low-latency connections). Simultaneously, the three traditional scenarios will struggle with IoT requirements as they grow. For example, IIoT requires both massive connections and significant UL data rates. Figure 1 illustrates several industrial applications in which the extensive and efficient deployment depends closely on the continuous evolution of 5G networks.

In addition to the existing eMBB, URLLC, and mMTC, 5G-Advanced will cover the following three new scenarios that require a major upgrade of the 5G network.

• UL centric broadband communication (UCBC), as a hybrid between eMBB and mMTC, will provide more UL-centric communication for applications that require support for massive connection density simultaneously with high upload bandwidth such as HD/UHD video uploading. UCBC can also significantly improve user experience with mobile phones in indoor scenarios with its larger UL capacity and deeper coverage. It will assist the acceleration of the digital transformation of the entire society, where the demand for UL business in ToC (e.g., interactive XR and multimedia social), ToB (e.g., remote operation and machine vision), IoT fields (e.g., video surveillance in smart cities) has increased significantly.

• Real-time broadband communication (RTBC), as an amalgam of eMBB and URLLC, will deliver ultra-wide bandwidth and low communication latency, which can be adopted in areas such as advanced XR experience and even holographic display services. Eventually, RTBC will facilitate the integration of the physical and digital worlds. By 2030, over 50% individual activities will be conducted in a digital space that is integrated with the physical world to an unprecedented extent.

• Harmonized communication and sensing (HCS), as the most novel use case, will provide both communication and sensing functions with 5G to enable services such as localization and autonomous driving via vehicle to everything (V2X). Through software and hardware upgrades, sensing capabilities can be provided using widely deployed 5G networks, which have inherent networking advantages, strong coverage, and large-scale antenna arrays. This will ensure that 5G network can be used in more industries in

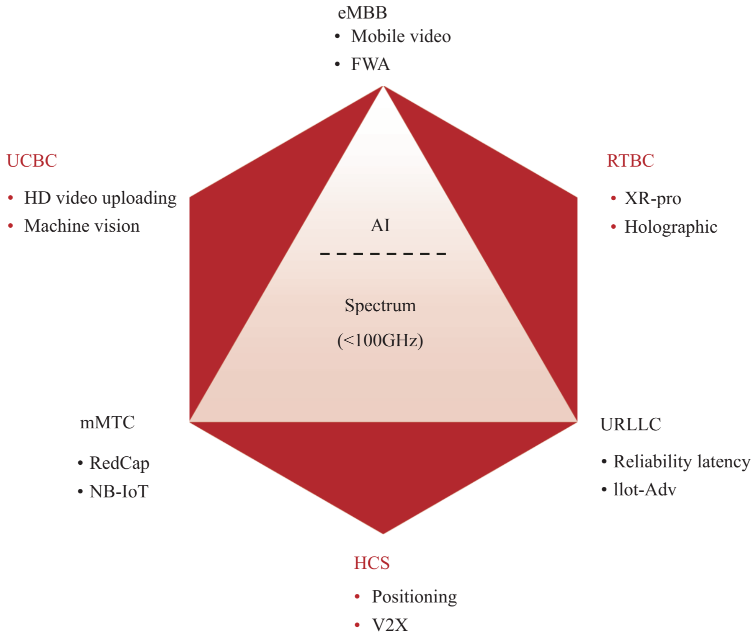

Figure 2 (Color online) Application scenarios and typical applications of 5G-Advanced.

the future.

Moreover, if 5G is seen as the Internet of everything (IoE) that connects the device-based IoT and human input elements, then 5G-Advanced is the intelligent IoE that enables us to build a fully interconnected world. Thus, 5G-Advanced would also require more sub-100 GHz spectrum to meet the requirements in the next decade. Full-band UL, downlink (DL) decoupling, and full-band carrier aggregation (CA) on-demand should use the spectrum more efficiently [21]. Furthermore, 5G-Advanced must be fully integrated with artificial intelligence (AI) to address the complexities caused by more frequency bands, device types, services, and customers than any previous generation. Figure 2 shows that the previous triangle of 5G is upgraded to a richer hexagon where its six applications scenarios (such three old scenarios defined in 5G) and representative applications under each scenario are clearly illustrated.

2.3 Target key performance indicators

In 2015, 5G KPIs were originally set by IMT-2020, which include 10 to 100 times of the 4G data rate for eMBB, 1 ms user plane latency in URLLC, and 10-year battery life for mMTC [22]. The evolution of 5G to 5G-Advanced must meet the increasing demands for immersive experiences such as XR and holograms and the diverse and complex IoT requirements such as machine vision and V2X. In these scenarios, Huawei predicts that 5G-Advanced will surpass 5G in several ways. For example, from 5G onwards, 5G-Advanced will reduce the cost-per-bit of eMBB by ten times, provide higher UL capabilities through UCBC, meet 80% of ToB scenarios, deliver Gbps DL rates at a low latency of 5 ms to double the number of XR users by RTBC and implement high-precision indoor positioning in all scenarios with low-power centimeter-level positioning and wide-area high-resolution sensing, thus enhancing the safety of self-driving cars and elder care in specific areas.

Based on what we discussed here and what was analyzed in Subsections 2.1 and 2.2, we propose the following main representative KPIs for 5G-Advanced [10, 18, 20].

- A 10-fold increase in user-experienced DL data rate for eMBB, e.g., from 100 Mbps to 1 Gbps.
- A 10-fold increase in user-experienced UL data rate for UCBC, e.g., from 50 to 500 Mbps.
- A 10-fold increase in bandwidth given a certain level of latency and reliability for RTBC, e.g., enabling XR services to achieve Gbps at 5 ms experience at any time and place.
- A 10-fold improvement in positioning accuracy from decimeter-level to centimeter-level for HCS.
- A 10-fold increase in connection density from one to ten million devices per square km.
- An end-to-end (E2E) latency with given reliability at a level from 5 to 4 ms at 99.999% and 99.9999%, respectively, for URLLC.

Table 1 Top 10 technology directions for 5G-Advanced

Domain	Direction	Scenarios	Subsection
Transmission	Multi-antenna transmission enhancement	eMBB UCBC	3.1
	Multi-antenna transmission evolution	eMBB	3.2
	XR delivery optimization	RTBC	3.3
	IIoT capability improvement	URLLC mMTC	3.4
	AI-enabled air interface	AI	3.5
Spectrum	Harmonized communication and sensing	HCS	4.1
	High frequency capability enhancement	UCBC HCS	4.2
	Spectrum value maximization	UCBC eMBB	4.3
Network	Network energy efficiency optimization	All	5.1
	Advanced wireless network architecture	All	5.2

These KPIs are a response from Huawei for future business forecasts and market investigations. Furthermore, they are a type of guidance and commitment to the 5G industry. Note that the KPIs will be achieved not only by the technologies introduced in the rest of this study but also by additional implementation-related technologies/algorithms, system implementation, and optimization.

2.4 Potential technology directions

In 1932, Guglielmo Marconi, the father of wireless communications, said that "It is dangerous to put limits on wireless". Indeed, both academia and industry have proposed several new emerging technological trends for 5G-Advanced and beyond. To reflect the aforementioned driving forces and to make the content concise, this study focuses on high-performance technologies, backward compatibility, and strong standard influence. However, some technologies are beyond the scope of this study. Technologies designed for new air interfaces such as waveform, coding, and THz, are incompatible with 5G and are often considered 6G-oriented technologies. Implementation-oriented technologies and proprietary system optimization technologies are excluded because of a lack of standard impacts.

Table 1 shows how we classified 5G-Advanced-oriented technologies into ten directions in three domains; it also shows the main application scenarios and the corresponding sections for each technological direction. Further, we will expand on the details of each direction in the rest of this study.

3 Technological directions in the transmission domain

3.1 Multi-antenna transmission enhancement

Here, 5G multi-antenna transmission technologies can be further enhanced along the direction of higher resolution in the frequency domain by applying finer-granularity precoding in the spatial domain using joint transceiving.

3.1.1 *High-resolution MIMO*

With MIMO being widely deployed in practical application scenarios, massive MIMO is gradually becoming an essential technology for providing high spectrum efficiency because of its high resolution. Academia and industries have been working to improve spatial resolution for nearly a decade. Using frequency division duplexing (FDD) as an example, the first release of the 5G NR specification (Rel-15) supports high-resolution codebooks in the spatial domain (SD), which is called Type-II codebook. The Type-II codebook is intended to facilitate quantized channel eigenvector feedback, which is notably beneficial for spatially multiplexing data transmission to multiple users, known as multi-user MIMO (MU-MIMO) [23]. In Rel-16, the eType-II codebook takes advantage of high channel correlation in the frequency domain (FD) to compress and jointly quantize channels from all sub-bands. Moreover, the overhead, which is saved by FD compression, can be used to improve the resolution; for example, more spatial beams can be configured, and higher resolution with more quantization bits can be used for amplitude and phase coefficient quantization [24]. In Rel-17, angle-delay reciprocity between DL and UL [25] is used to improve the resolution and acquisition of channel state information (CSI). Thus, the SD/FD compression operation inherent in Rel-16 eType-II codebook can be shifted to the gNB (i.e., 5G base station gNodeB) to further improve the resolution [26].

Although high-resolution MIMO has undergone several developments and significantly improved the spectrum, further enhancement of high-resolution MIMO for both FDD and time division duplexing (TDD) is beneficial for providing higher spectrum efficiency for the upcoming scenarios; however, the penetration of 5G users is rapidly increasing and new applications for high data rate requirements are being developed.

Initially, the best sparsity for transforming domain is left and right eigenvectors using singular value decomposition (SVD). However, because of limited implementation complexity and CSI feedback overhead, SVD cannot be supported for each reporting instant. In Rel-15 and 16, a discrete Fourier transform (DFT) basis is used for the spatial and frequency domain. In Rel-17, the statistics eigenvector of the UL spatial-frequency covariance matrix is considered the basis for the spatial and frequency domain. Although it can be proved that the reciprocity of DL and UL spatial-frequency covariance matrix exists if statistical time is long enough, there is still a gap in their covariance matrix in practice with limited statistical time. Thus, statistical eigenvectors of the DL spatial-frequency covariance matrix can be considered a basis for further improving the CSI feedback accuracy.

Furthermore, the granularity of precoding is restricted as wideband, which are four or two resource blocks (RBs) in Rel-15–Rel-17. However, in some practical deployments, the channels have large frequency selectivity. The angle power profile of these channels shows that the beam varies within one precoding resource block group that comprises four or two RBs. Thus, finer precoding granularity can be considered to further improve DL MIMO performance in highly frequency selective channels. Note that more advanced channel estimation is required at the user equipment side to fully exploit such benefits.

Furthermore, CSI aging would introduce performance loss, particularly in the mobility case because of high Doppler. CSI prediction is a potential solution for ensuring performance in the mobility case. The Doppler or Doppler-delay spectrum-related information can be used to characterize the channel variations over time [27], which can be measured on multi-instant CSI reference signal (CSI-RS) resources. Thus, a possible method for CSI prediction is a joint CSI feedback from multi-instant measurements.

Additionally, the FDD system depends heavily on CSI-RS, whereas the TDD system relies on sounding reference signals (SRS) to acquire the channel information. As mentioned above, in Rel-17, FDD channel reciprocity is thoroughly investigated in the CSI feedback, which provides significant performance gain compared with conventional CSI design. Meanwhile, UE (user equipment) or UE group-specific CSI-RS is required for FDD CSI, based on Rel-17. Considering that the penetration of 5G users is quickly increasing, the CSI-RS overhead, particularly for periodic CSI-RS in such cases, could be further studied. For example, multiplexing CSI-RS ports on the time delay domain can also be considered to reduce CSI-RS overhead. Similarly, the SRS capacity should be enhanced for the TDD system. Rel-17 introduced partial SRS for SRS capacity enhancement, in which it is transmitted only on some but not all physical resource blocks (PRBs) of the hopping band [28]. Based on Rel-17 partial SRS, two and four times SRS capacity enhancement can be expected. To further enhance SRS capacity in the case of dense users, a new SRS design must be considered by reducing the SRS overhead based on the sparse characteristics of multi-path channels or producing additional orthogonal SRSs with zero auto/cross-correlation zones, which can guarantee the channel estimation accuracy.

3.1.2 *Multi-cell coherent joint transceiver*

For the eMBB scenario in 5G-Advanced, the traffic load will significantly increase with additional transmission data streams and interference in the network. Interference suppression for intra-cell and inter-cell will be a critical problem for further system performance enhancement. The evolution of network architecture also provides the possibility of large-scale multi-cell coordination. With the increasing proportion of cloud radio access network (C-RAN) deployment globally, user-centric no-cell (UCNC) can be better achieved in 5G-Advanced, resulting in a near-uniform experience for users throughout the network.

In the C-RAN scenario, distributed MIMO (D-MIMO) or coherent joint transmission (CJT), which combines multiple MIMO antenna arrays into a much larger one, is an efficient way to improve the system spectral efficiency and user experience, particularly for cell-edge users. Furthermore, effective cancelation or suppression of the interference from neighboring cells is crucial when trying to improve performance. Precise inter-cell interference suppression requires interaction between coordination cells, which is possible for C-RAN or intra-site scenarios. The coordination within a large-scale cell set, which includes tens of cells, can be achieved in a centralized or distributed approach to improve the system performance significantly.

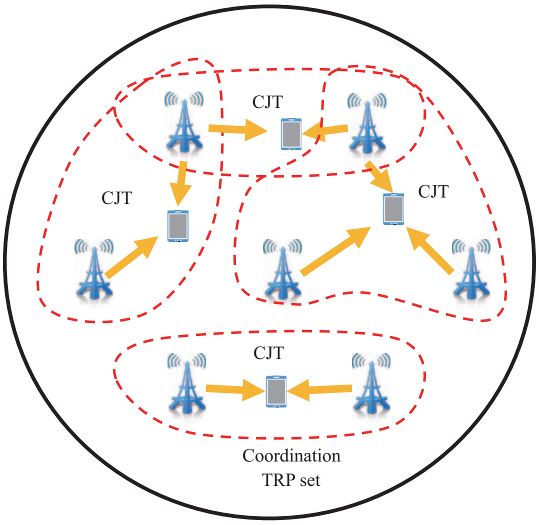

Figure 3 (Color online) An illustration of a large-scale multi-cell coordination.

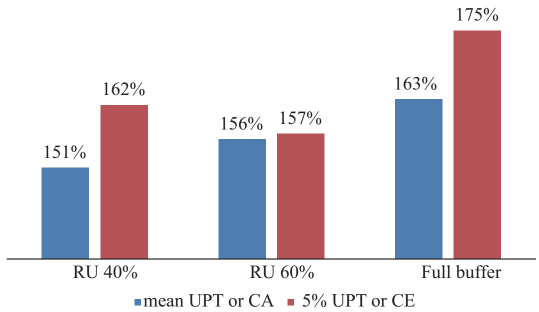

Figure 4 (Color online) Performance gain of high-resolution CJT CSI over Rel-17 for FDD CJT. UPT: user perceived throughput, RU: resource utility, CA: cell average, CE: cell edge.

Figure 3 illustrates a typical large-scale multi-cell coordination. The coordination cell set (cell set inside the black solid line) contains several cells. For different UEs, the transmission cell sets (cell sets inside the red dashed lines) for CJT are different. In this coordination scenario, the interference suppression can be conducted within the entire coordination cell set, and the joint transmission can be UE-centric.

For CJT with more cells and precise interference suppression, the precision of channel information for multiple cells is critical for each UE. If the channel information is from UE CSI feedback, each UE must measure the channel of the cells at least, including the coherent transmission cells. Existing 5G CSI measurement designs for single-cell or non-coherent joint transmission and high-resolution CSI feedback designed for CJT are necessary for 5G-Advanced. Channel measurement of more cells increases DL-RS resource and UL feedback overhead. However, achieving better multi-cell CSI acquisition with limited overhead is a critical problem. Figure 4 shows an exemplary CJT performance evaluation, where over 50% gain can be achieved via high-resolution CJT CSI compared with that obtained via Rel-17 CSI.

SRS is used for channel acquisition for TDD CJT. Because channels from each UE to multiple cells are required, the inter-cell interference of SRS should be considered for multi-cell cooperation. SRS enhancement for measuring UE channel using multiple cells with less interference is also a critical problem to investigate in 5G-Advanced. The following are several potential research directions.

- Advanced SRS design pattern or sequence for better performance in the inter-cell interference case.
- Sequence allocation and power control of SRS for reducing the interference in multi-cell channel estimation.
- Enhanced SRS channel estimation algorithms for measuring channels in a lower signal-to-interference-plus-noise ratio (SINR), considering the non-orthogonality of SRS introduced by multi-cell channel measurement.

3.2 Multi-antenna transmission evolution

Besides technological enhancements, more advanced and promising multi-antenna technology evolution for 5G-Advanced or even 6G has received increasing attention in both academia and industry. This subsection introduces three of the most prominent evolution technologies, such as extremely large aperture array (ELAA), reconfigurable massive MIMO, and intelligent reflecting surface (IRS).

3.2.1 *Extremely large aperture array*

With the successful deployment of massive MIMO equipped with 64 fully digital transceiver chains, we can expect numerous antennas to serve a set of users in the future. However, it is challenging to deploy a compact, co-located massive MIMO array with a significant number of antennas traditionally because of

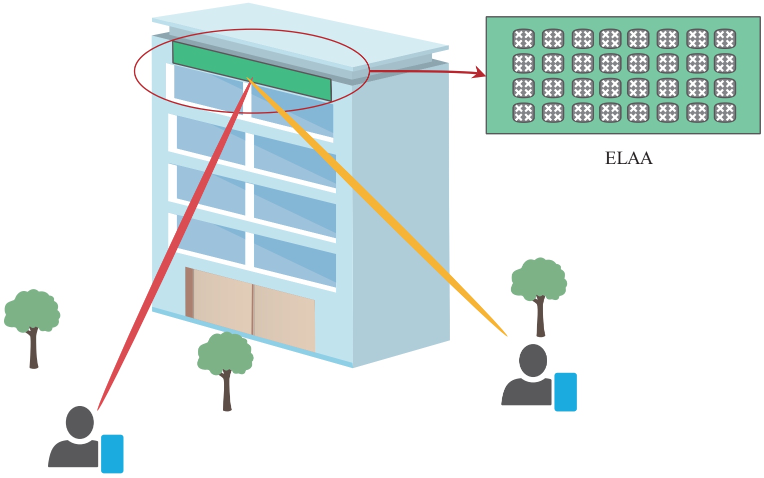

Figure 5 (Color online) An illustration of ELAA.

practical limits such as panel size, weight, and wind load. Thus, we must seek a new antenna deployment strategy. Rather than gathering all antennas in one panel, they can be distributed over a substantially larger area, which is referred to as ELAA [29]. Figure 5 shows an example where the antennas are deployed along the wall of tall buildings.

Since ELAA is a huge MIMO deployment with numerous antennas, the ELAA channel shares similar features observed in massive MIMO, such as channel hardening [30], favorable propagation [30,31], and low eigenvalue spread [32,33]. Moreover, users may be in the array's near-field because of the large aperture of ELAA. Thus, the traditional planar wavefront assumption is invalid, and the spherical wavefront should be considered instead [34]. Another essential property of the ELAA channel is spatial non-stationary. Specifically, different portions of the array can see different scatters and receive different power levels [35].

For the ELAA system, because of the large number of antennas, the complexity and fronthaul data rate of baseband processing increase significantly. Thus, the distributed baseband processing algorithm and architectures with low complexity and fronthaul data rate should be investigated. To develop distributed schemes, antennas can be grouped in clusters, and several schemes have been proposed based on this assumption. For example, alternating direction method of multipliers-based equalizer and precoder were proposed in [36]; however, this suffers from data transfer latency because of the iterative exchange of information among antenna clusters. To address this latency issue, feed-forward architectures for UL detection [37] and DL precoding [38] are proposed. The daisy chain was proposed [39] and an optimal minimum mean square error UL detection under colored noise was proposed for the architecture [40]. To exploit the spatial non-stationary of ELAA channel, certain low-complexity UL detection schemes have been proposed [41]. Although several studies have investigated the distributed baseband processing algorithm, there are still some issues that must be addressed. First, for UL detection design, colored noise caused by interference is rarely considered. Although colored noise was considered in [40], the authors assumed that the noise is independent among clusters. In reality, the antenna clusters are close to each other; thus, such an independent assumption is invalid. Therefore, the equalizer under colored noise is currently an open problem. Second, although several schemes have been proposed to reduce the fronthaul data rate, the data rate is still high in the case of high data streams, posing a great challenge for the fronthaul design; for example, only parts of the array contribute to user performance because of the spatial non-stationary of ELAA channel, which implies that the fronthaul requirement can be reduced without severely reducing the performance. Furthermore, the fronthaul data rate for only transferring the quadrature amplitude modulation (QAM) symbol sequences will be over 80 Gbps in a system with 100 MHz bandwidth and 128 data streams, and there is no room to compress such QAM symbol sequences further without any loss. Thus, a new transmission or equalization strategy is required, which remains an open problem.

Furthermore, channel modeling is another open problem for the ELAA system. The widely used 3GPP NR channel model [42] fails to capture the near-field effect and spatial non-stationary of the ELAA channel. However, the COST 2100 channel model [43] can capture the physical properties of the ELAA channel, but it is inefficient to model the channel in a scenario with multiple UEs and BSs. Thus,

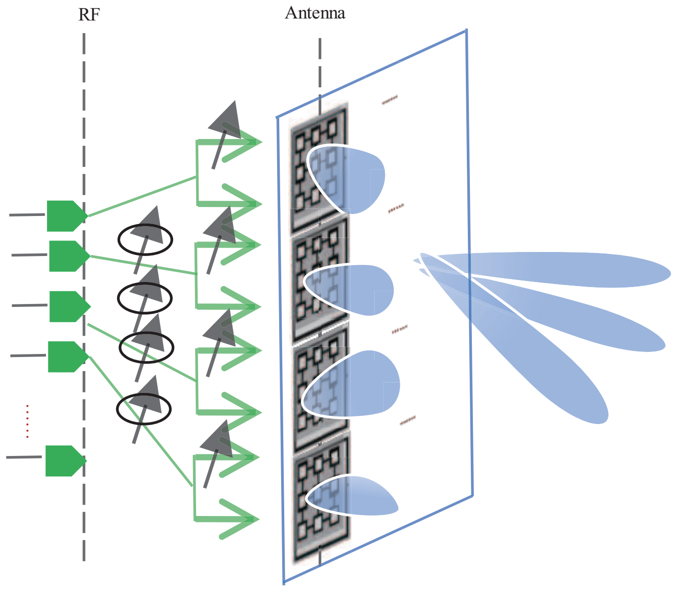

Figure 6 (Color online) A realizable form of reconfigurable massive MIMO.

measurement campaigns and improved channel modeling that capture physical features are required.

Note that besides the ELAA system discussed above where all the antennas are deployed at the same site, there are also more general concepts under the ELAA umbrella in a large sense [29], such as cell-free massive MIMO and D-MIMO where numerous antennas are deployed in different locations far from each other. Therefore, the channels and interference of different locations are considered statistically independent of each other [44].

3.2.2 *Reconfigurable massive MIMO*

Reconfigurable massive MIMO is proposed as another possible solution for breaking the limit of MIMO capacity with a given antenna size [29,45]. Thus, the related air interface, algorithm, hardware, and even the fundamental theory must be studied. Figure 6 shows that a possible realizable form of reconfigurable massive MIMO is to jointly design the digital, RF, and antenna domains to achieve the maximum capacity.

The related key issues are as follows:

• From the aspect of hardware. A high-performance reconfigurable antenna design (directivity pattern reconfigurable) is required.

• From the aspect of air interface. Since multi-antenna mode must be measured, low-overhead and high-precision measurement schemes are crucial for antenna mode and CSI, e.g., joint estimation scheme for antenna mode and CSI [46].

• From the aspect of signal processing. Based on reconfigurable antennas, as new dimensions are introduced using antenna modes compared with traditional massive MIMO or hybrid beamforming (HBF), new signal processing methods, such as media-based and polarization modulation, should be studied to fully use dimension and to significantly improve the spectral efficiency of the system.

For example, because more dimensions are jointly designed, the complexity is an important factor in signal processing; thus, a two-stage design [47] or Thompson sampling [48] have been proposed to reduce the complexity of precoding algorithms. Moreover, the multi-cell cooperation with the reconfigurable massive MIMO could be different from that with massive MIMO because the reconfigurable antenna will introduce more challenges for interference suppression. A possible solution is to use the randomized two-timescale hybrid precoding to achieve a better tradeoff between sum throughput and fairness while mitigating inter-cell interference with reduced CSI overhead and RF chains [49].

Another exciting direction introduced by the reconfigurable massive MIMO is that new communication theories, such as electromagnetic information theory [50], may bring new opportunities to communication theory. Electromagnetic information theory is the combination of information and electromagnetic theories, which models the physical (electromagnetic) communication transmission model, transmission rate, and spatial degree of freedom, and the physical electromagnetic characteristic dimensions are established.

Table 2 Potential commercial use cases of IRS

Typical scenarios	1st phase	2nd phase
Outdoor	Outdoor coverage: IRSs are deployed to improve the coverage, e.g., high frequency or coverage hole.	Throughputs & user experience: multiple IRS are deployed in multi-cell networks to enable high throughputs of MU scenario and better user experience.
Indoor	Indoor coverage: for indoor scenario, e.g., home or industry, IRSs can be deployed to improve the coverage of signals.	URLLC & throughputs: for industry scenario, e.g., industry 4.0, multiple IRS can be deployed to further ensure the URLLC and improve throughputs.

It is expected to provide a new evaluation and guidance for the performance and algorithm design to approach the ultimate performance of finite physical antenna apertures, such as near-field performance analysis and optimization of extended physical antenna apertures, polarization modulation that might break the upper bound of multi-antenna transmission, and coupling assisted precoding design. Related key technologies are as follows:

• Basic theory and modeling [50, 51].

• New antenna design, such as design based on mode domain that respond backward to match the wireless environment [52] and environment-adaptive metamaterial antenna [53].

• Joint system and air interface design to achieve a higher capacity such as polarization modulation and demodulation [54].

• Capacity analysis and optimization such as coupling-aware precoding [55] and intelligent optimization of near-field electromagnetic environment [56].

3.2.3 *Intelligent reflecting surface*

IRS, or also called as reconfigurable intelligent surfaces, is a new outcome across wireless communication and material science [57, 58], with potential use cases such as hardware, key algorithms, and network deployment.

However, several key challenges must be studied before IRS is widely deployed to achieve commercial success; these challenges are as follows:

• Hardware architecture design of IRS such as power supply, considering the practical deployment environment and the price-performance ratio.

• The product size of IRS is restricted by the deployment method, e.g., on the wall or the top of a roof or a pole, indoor/outdoor, and via the installation specifications such as wind resistance.

• The dynamic level of IRS should consider the balance between performance and price. Moreover, it will directly influence the related air interface design, e.g., CSI measurement and signaling mechanisms.

• Related key algorithms for the physical layer of IRS such as precoding, channel estimation, multi-IRS, and multi-cell cooperation.

It should be clarified that the entire IRS design will be directly influenced by the most typical and commercial use cases. Based on the development of technology and hardware requirements from our perspective, the typical commercial use cases of IRS could be divided into two phases: the first phase for coverage satisfies the basic requirements, and the second phase is enhanced to achieve high throughputs and/or URLLC requirements, as shown in Table 2.

From the perspectives of air interface, algorithms, and hardware architecture, key technologies and challenges for IRS can be summarized as follows.

• To achieve the highest performance gain and benefit from the easy IRS deployment, most researchers consider the wireless link between the BS and IRS. However, obtaining the accurate CSI of IRS and controlling IRS in time with low air interface overhead is extremely challenging. Therefore, advanced air interfaces and algorithms should be designed for IRS. For the feedback, a possible solution is to design an adaptive bit partitioning codebook to feedback the CSI of IRS [59], which could also be combined with UL SRS to improve the CSI accuracy. For the MU scenario, the inner common properties between users' channels can also be used for channel estimation, which can reduce the RS overhead [60]. Moreover, for a multi-cell scenario deployed with multi-IRSs, the distributed channel estimation and precoding algorithm of IRS should be designed to achieve high network efficiency with low complexity [61–63]. Note that the mechanism and air interface design should be compatible with the evolution of multi-cell cooperation such as UCNC and UCBC. Furthermore, air interface design should consider different hardware IRS implementations such as with or without power amplifier (PA) because the PA may disrupt the channel

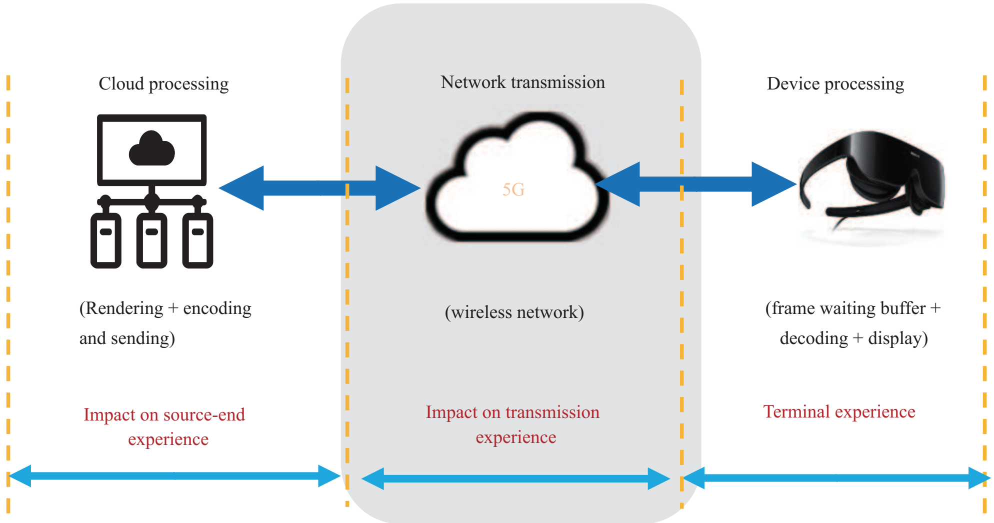

Figure 7 (Color online) Three parts related to XR service quality.

reciprocity of UL and DL channels. Moreover, advanced precoding schemes in IRS such as symbol-level-precoding (SLP) [64] can be further considered for higher performance.

• The hardware IRS architecture requires a good balance between several practical factors, such as supply conditions at BSs, deployment type of IRS, and price-performance ratio. Solar energy is a possible power supply mode that benefits from easy deployment. Specifically, the solar panel can be integrated with the IRS, or it can be separately deployed from the IRS. Another new IRS deployment is to further consider its transmissive properties such as intelligent Omni-surface (IOS) to extend its application scenarios [65]. Unlike the reflective properties of widely studied IRSs that restrict the service coverage to only one side of the surface, IOS can provide service coverage to mobile users in a reflective and transmissive method.

In conclusion, IRS is a promising technology in 5G-Advanced because of its unique advantages, such as easy deployment and changing the wireless channel. However, compared with other counterparts such as small cells and repeaters, whether IRS can achieve the final commercial success requires comprehensive studies.

3.3 XR delivery optimization

The most critical problem for high-interactive broadband communication represented by XR services is to develop unified quality evaluation criteria to guide transmission optimization and even network deployment.

3.3.1 *XR quality evaluation*

Depending on the traffic requirements, different services may have different KPIs. For example, eMBB services focus on peak data rate and throughput. URLLC services, such as factory automation, transportation, and power distribution, have clear requirements for reliability and latency. However, for XR services, quality of experience (QoE) is a crucial metric related to the end-user perception of the service. The features of interaction and immersion for the XR services make them more perceptive than traditional videos. Hence, existing KPIs such as throughput, reliability, and latency cannot directly reflect the user experience in XR services. For example, it may be unreasonable to determine whether an XR UE is satisfied just based on a particular packet error rate (PER) or packet delay budget (PDB). However, different PERs or PDBs can result in different user experiences. A lower PER or PDB can provide better XR quality, thus a better user experience and vice versa. Therefore, a new subjective performance metric can be considered to meet the characteristics of XR.

However, E2E user experience in XR service is influenced by these three parts: XR source, network transmission, and XR terminal parts [66]. Figure 7 shows an example of these parts. The current E2E metric cannot determine the part that causes bad XR quality, which complicates troubleshooting. Thus, a new KPI that can reflect the impact of network transmission on XR service quality can be considered to meet the characteristics of XR videos. For convenience, such a desired KPI is called XR quality index (XQI) in this study. The following benefits can be achieved using XQI in RAN:

• Network transmission impact on user experience can be better evaluated using XQI.

• XR measurable performance in operators' networks can be obtained and used for network construction planning and operation optimization.

The frame quality modeling could be used to define XQI (see (1)). Suppose that when a frame is lost, it can be replaced by copying the most recent decodable frame, thus resulting in a stalling artifact. Let us build a model to calculate the quality of each stalling frame. During a scoring period, the quality of the n-th frame is modeled as follows [67]:

$$D_{f(n)} = \begin{cases} 100, & n < i, \\ \max\left(v1 \times \ln\left(v2 \times \dfrac{n-i+1}{f_r} + v3\right) + v4\right), & n \geqslant i, \end{cases} \tag{1}$$

where i is the index of the first lost frame, f_r is the frame rate, $v1$, $v2$, $v3$, and $v4$ are fitting coefficients, which equal -3.6390, 55.9409, -0.9320, and 13.9274, respectively. Furthermore, the XQI in a scoring period can be calculated as follows:

$$\mathrm{XQI} = S_{\mathrm{el}} \frac{\sum D_{f(n)}}{N}, \tag{2}$$

where S_{el} is the average frame quality after source compression (e.g., $S_{\mathrm{el}} = 70$) and N is the total number of video frames in the scoring period.

3.3.2 *XR capacity improvement*

Although XR services are quite promising, providing XR services over 5G networks remains extremely challenging because of these characteristics: high average data rate, stringent latency requirement, and relatively high-reliability requirements. For example, according to Clause 6.2.4.2 in TR 26.928 [68], the increasing resolution and frame rate requirements for the XR video will increase the bit rate to 1 Gbps. A motion-to-photon latency of 20 ms (including media rendering, encoding, network delivery, and decoding) is required for strong interactive XR applications and cloud rendering to meet the immersive limits. For reliability, a packet error rate less than $10^{-3}/10^{-4}$ for UL/DL transmission is required to guarantee a satisfactory QoE [69]. According to the initial 3GPP evaluation results [70], only few XR users can be supported in a cell, which is far from commercial use. This is mainly due to the 5G air interface not being particularly optimized for the specific requirements of XR applications. Therefore, several potential techniques for XR capacity enhancement are presented as follows.

(1) Layered QoS. For a given XR application, there could be multiple data streams with different traffic characteristics and QoS requirements in both DL and UL. However, in the current 5G QoS framework, multiple data streams belonging to the same XR service are transmitted over the same QoS flow. Consequently, these data streams cannot be identified, and the network will always treat them equally, resulting in over-protection for non-important packets and potential waste of radio resources, which causes low capacity. Even if the application layer may already split the important and unimportant parts into different QoS flows, the 5G network is unsure whether these QoS flows are associated and will process packets over multiple QoS flows independently. This may result in non-synchronization of packet transmission from the same source and the failure of (de) prioritizing corresponding packets. Thus, service quality cannot be guaranteed in this scenario.

To accommodate the characteristics of multiple data streams in XR services, a layered QoS mechanism with multiple QoS flows can be considered, with two data streams, as illustrated in Figure 8. The data stream/flow with a higher QoS requirement is referred to as basic flow, whereas the other with a lower QoS requirement is the enhancement flow.

However, the existing mechanism also supports multiplexing multiple streams of an XR source into a single QoS flow, where a layered QoS mechanism can still be considered, as shown in Figure 9. Here, some E2E mechanism for distinguishing basic and enhancement parts based on enhanced QoS information is required.

In RAN, prioritizing the transmission of the more important stream is essential for expanding capacity. Based on the initial evaluation results [71], under typical system evaluation, such a scheme can improve the XR capacity by approximately 61% compared with the existing mechanisms such as proportional fairness (PF) scheduler.

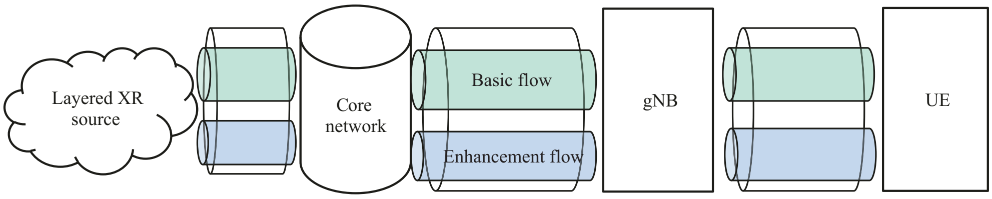

Figure 8 (Color online) Layered QoS mechanism with multiple QoS flows.

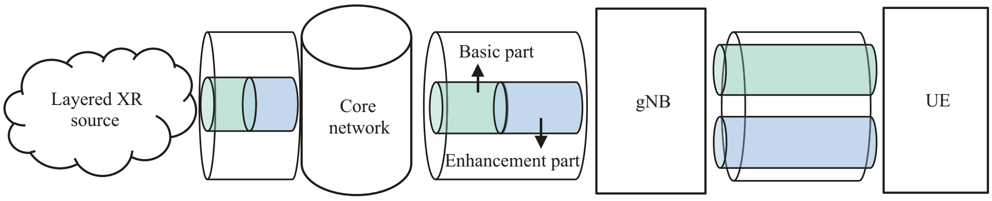

Figure 9 (Color online) Layered QoS mechanism with single QoS flow.

(2) Integrated transmission. From a network transmission perspective, each video frame in XR services may be segmented into one or multiple correlated Internet protocols or packet data convergence protocol (PDCP) packets. Without the tiling or slicing approach, a video frame can be properly decoded and reconstructed only when all its associated packets have been correctly received. However, existing RANs are unaware of the correlations between packets in XR applications. If one or more packets associated with a video frame are lost, the video frame cannot be decoded even if the subsequent packets are successfully transmitted, resulting in a waste of radio resources.

To support the frame integrity, coordination between the application and the 5G core network for identifying packets that belong to a video frame is beneficial for satisfying the XR service requirement. This information about grouped packets can be subsequently communicated to the RAN. RAN can benefit from this frame-level integrated transmission to enable efficient radio resource management. For XR services, the network capacity is defined as the maximum number of users in each cell, and at least $X\%$ (e.g., $X = 90$) of users are satisfied. For the definition of network capacity, if a user's frames are successfully transmitted over $Y\%$ (e.g., $Y = 99$) within a given frame delay budget (FDB), the user is considered a satisfied user. Thus, we can investigate the network capacity maximization problem, which maximizes the number of satisfied users under the data rate, reliability, and delay requirements of each user with a given bandwidth per BS. This problem can be expressed as follows [72]:

$$\begin{aligned} &\max_{x} |A| \\ &\text{s.t. } E_f\left[\mathrm{FSR}_{k,f}\left(x, \mathrm{FDB}_k\right)\right] \geqslant e_k, \quad \forall k \in A, \\ &\qquad E_f\left[R_{k,f}\left(x, \mathrm{FDB}_k\right)\right] \geqslant r_k, \quad \forall k \in A, \\ &\qquad A \subseteq K, \end{aligned} \tag{3}$$

where K and A are the set of candidate users and the set of satisfied users, respectively. $\mathrm{FSR}_{k,f}(x, \mathrm{FDB}_k)$ and $R_{k,f}(x, \mathrm{FDB}_k)$ are the achievable frame success rate and the achievable data rate of the f-th frame of the k-th user under the resource allocation scheme x and the delay constraint FDB_k, respectively. e_k and r_k indicate the target frame success rate and the target data rate of the k-th user, respectively. E_f is the expectation operation over all frames.

System-level evaluation results are provided [72] to verify the effectiveness of the frame-level integrated transmission scheme. Figure 10 illustrates a performance comparison with an existing mechanism in terms of user satisfaction rate. Here, if a user's frame is successfully transmitted over $Y = 99\%$ within the given FDB, the user is considered a satisfied user. The user satisfaction rate is the ratio of the number of satisfied users to the total number of users in the network. The results show that the frame-level integrated transmission scheme (denoted as integrity) achieves a much higher user satisfaction rate than the PF scheduler. The proposed scheme can increase the network capacity from four to ten users, resulting in a 150% improvement compared with the PF scheduler. The network capacity is defined as the maximum number of users in each cell, with at least $X = 90\%$ of users being satisfied.

(3) Network coding for XR. Since the theory of network coding was proposed in 2000, it has been widely applied to multiple fields such as network transmission and distributed storage. With the broad

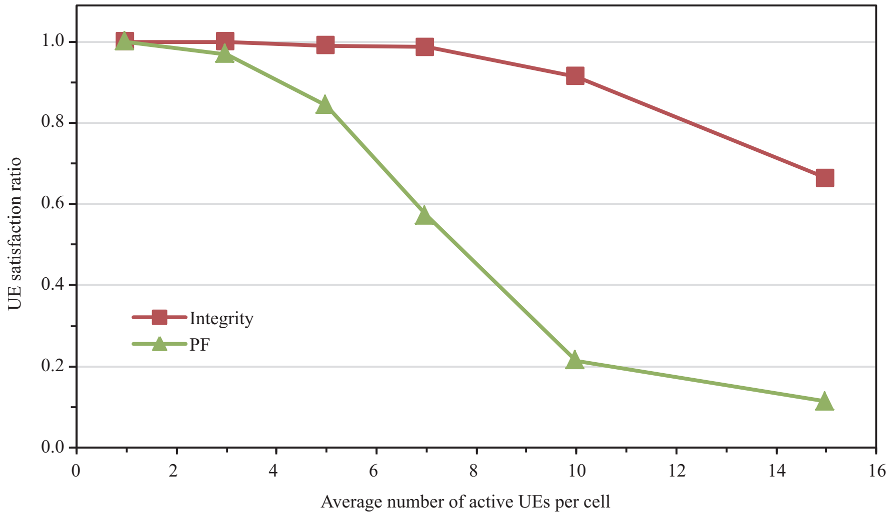

Figure 10 (Color online) User satisfaction ratio versus the average number of active users per cell.

application of real-time multimedia services and the emergence of diversified scenarios of 5G low-latency and large-bandwidth services in recent years, network coding technology has ushered in a new wave of research. A key point of network coding is the study of code type. Similar to traditional channel coding, network coding can also be divided into two categories: packet and convolutional network coding. Typical examples of packet network coding are Reed-Solomon code [73], random linear network coding [74], and raptor code [75]. The convolutional network coding introduces a certain overlap between different groups of original packets; it has a semi-infinite generator matrix such as MDS convolutional code [76], m-MDS code [77], and streaming code [78].

For the XR service, network coding has the following potential technical directions.

• Integrity protection. By adopting systematic network coding, a small portion of code packets are generated based on the original packets from the same frame, which can effectively prevent frame loss caused by channel fading or burst interference.

• Unequal protection. If multiple QoS flows are supported over the RAN air interface, different coding schemes can be applied to different QoS flows; for example, high-priority QoS flow is encoded to more parity packets, whereas low-priority QoS flow is encoded to fewer parity packets.

• Anti-Jitter. In jitter-sensitive scenarios, a conservative block size of block code should be configured to facilitate parity packets with a given latency bound, which inevitably causes some spectral efficiency loss. However, convolutional network coding is more robust to jitter because the receiver side can determine the decoding depth adaptively based on the real-time delay budget, resulting in the best spectral efficiency possible.

Network coding has the following advantages:

• Lower latency than the retransmission strategy.

• Higher spectral efficiency than the repetition strategy.

• Less feedback overhead than per packet feedback.

As a potential basic 5G-Advanced technology, network coding can also be applied in several scenarios such as URLLC, integrated access and backhaul (IAB), dual connectivity (DC), multicast, and V2X. Thus, a common network coding protocol stack architecture should further be designed.

3.4 IIoT capability improvement

IIoT capability refers not only to the data rate for eMBB but also to the latency and reliability for URLLC and the massive connectivity for mMTC. Furthermore, IoT devices with ultra-low power consumption are extremely attractive in certain scenarios. We will address the above-mentioned four dimensions in detail below.

3.4.1 *UL capability*

For IIoT eMBB, as mentioned above, there is an increasing demand for UL capacity. Since both the power and number of antennas on the UE side are limited and cannot be increased unconstrained, maximizing MIMO capabilities with a limited number of UE antennas is a direction worthy of further consideration and research. Additionally, UE cooperation is a means for expanding the UE UL capability.

(1) Enhanced UL MIMO. Here, 5G NR introduced the multiple transmit/receive point (multi-TRP) operation, where a serving cell can schedule UE from two TRPs. This is particularly beneficial for the factory scenario with several TRPs deployed to serve numerous users along the assembly lines. Furthermore, multi-TRP UL joint processing is considered a potential means for improving the UL capacity. Thus, UL precoding, power control, and demodulation reference signal (DMRS) should be further enhanced to efficiently enable the multi-TRP UL joint processing.

• UL precoding. 5G supports two UL transmission schemes such as codebook- and non-codebook-based UL transmissions. Codebook-based UL transmission can only use UL wideband coarse codebook (only constitutes of ± 1 and $\pm j$). Furthermore, it cannot match irregular UE antenna shapes or patterns, such as UE antennas with non-uniform spacing. However, non-codebook-based UL transmission is based only on each UE's channel information without considering MU interference. Additionally, for both codebook- and non-codebook-based transmissions, the UL precoder is wideband, i.e., only one precoder is applied to the entire bandwidth, which can be enhanced via sub-band precoding. Furthermore, in some industrial applications, a powerful UE may be equipped with over two antennas (four or eight antennas). These antennas can be used to boost the UL capacity and improve the overall network capacity, as long as higher-resolution uplink precoding is provided. In our initial simulation, the sub-band precoding and higher-resolution precoder in an IIoT scenario show approximately 20% cell average throughput gain; however, this may lead to large system overhead. Thus, the UL precoding indication mechanism should be optimized. In summary, these two questions are required to enhance UL precoding:

(a) How to optimize UL MU precoding considering MU interference with achievable low complexity and low interaction overhead.

(b) How to indicate high-resolution sub-band UL precoding dynamically using low DL indication overhead.

• UL DMRS. Presently, up to 12 orthogonal antenna ports are specified in the current 5G specification; however, multi-TRP joint processing significantly increases the probability of over 12 concurrent UL layers in a local area. Non-orthogonal antenna ports can be configured to handle over 12 UL layers, but it degrades the accuracy of UL channel estimation because of relatively high cross-correlation among DMRS sequences [79]. In such scenarios, the maximum number of orthogonal or low correlation UL antenna ports should be increased. To address the DMRS port limitation issue, the following two types of enhancement directions must be taken:

(a) Introduce more orthogonal cover codes (OCC) without additional DMRS resources, such as length extension of the OCC for each CDM group. Note that the frequency OCC can be seen as a special case of multiplexing in the time delay or cyclic shift domain.

(b) Introduce low cross-correlation DMRS sequences to address the increased cross-correlation.

Our initial evaluation shows that the performance of approximately 24 layers with DMRS enhancement can achieve a 68% cell average gain compared with non-orthogonal DMRS of the current specification.

• UL power control. Additional UL layers and joint UL multi-TRP detection result in a higher requirement on UL power control. However, for UL MU-MIMO, a user may be paired with different users in different slots, and these dynamic user pairs can lead to large UL transmission power variation. The variation can be larger than the existing power control adjustment steps. Both closed- and open-loop power control can be enhanced to match this large power variation. However, the UL power control should be adopted for multi-TRP reception because of the different received power on each TRP. This implies that gNB optimizes the UE transmitting power based on the path loss of multiple TRPs involved in the joint processing instead of on the single path loss from only the serving TRP. Our initial evaluation shows that enhanced power control can provide approximately 20% cell edge performance gain.

(2) Network assisted UE cooperation. UE cooperation via sidelink (SL) device-to-device communications can be used to improve the overall UL experience for both capacity and coverage. For example, in a scenario where data rate requirements or coverage requirements cannot be met, a cooperative UE (CUE) can be introduced to operate cooperatively with the source UE (SUE). CUE and SUE can be connected via SL to form a virtual UE with a higher capability because of the aggregated transmit power

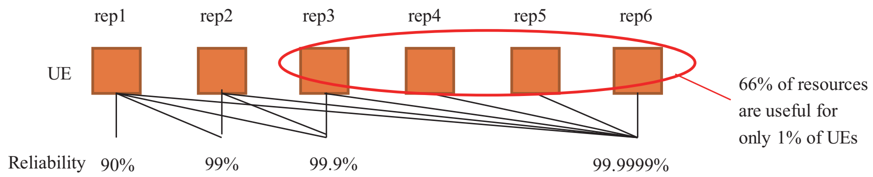

Figure 11 (Color online) Reliability increases as the number of received repetitions increases.

and aggregated distributed transmit antennas. When the SUE is out-of-coverage, the CUE acts as a relay to forward the SUE's UL data to the network. UE cooperation could be viewed as an extension of the UE relay. Additionally, CUE can be used as a diverse UE or a backup UE of SUE to ensure URLLC requirements, especially in IIoT applications. However, this type of UE cooperation should be under the supervision and management of the network; thus, network-assisted UE cooperation was proposed [80] as a promising technology in 5G-Advanced.

3.4.2 *URLLC capability*

The basic functionalities for URLLC were standardized in NR Rel-15, which includes features for achieving the targeted 1 ms latency and 1e−5 reliability. Further, physical (PHY)/higher layer enhancements targeting 1 ms latency and 1e−6 reliability and features such as inter/intra-UE service prioritization/multiplexing were specified or are being specified in NR Rel-16/Rel-17. However, for practical network deployment for supporting URLLC services, more realistic factors must be considered, such as the TDD UL/DL configuration that should be used considering the co-existence with eMBB services and how to improve the system efficiency while ensuring the stringent reliability requirement with inter-cell interference. Here, we propose several potential technical directions to be enhanced in 5G-Advanced.

(1) Retransmission enhancement. Retransmission is a crucial technology for reducing the delay and improving the reliability of URLLC services. For eMBB services, feedback-based retransmission is commonly used, which cannot meet the requirements of URLLC services with extremely low latency because of the time-consuming interaction. Thus, blind retransmission (aka. repetition) was introduced in 5G NR as a means of reducing delay and improving reliability by retransmitting before receiving feedback. Figure 11 illustrates that six repetitions (i.e., five retransmissions) can achieve 1e−6 reliability without waiting for feedback. However, high reliability provided via blind retransmission comes at the expense of spectral efficiency.

To resolve the conflict between spectral efficiency and transmission latency, several blind retransmission enhancement techniques have been proposed [81–85]. Blind retransmission enhancement under UL contention access is studied to improve spectral efficiency while ensuring reliability [81, 85]. A hybrid retransmission mechanism has been proposed [82, 84] where feedback retransmission and blind retransmission are combined. A blind retransmission mechanism based on mixed and non-contention resources was proposed [85], where some retransmission resources of multiple UEs are shared.

(2) URLLC & eMBB multiplexing. The coexistence of URLLC and eMBB services is also a critical problem because several types of services with different characteristics would be simultaneously delivered wirelessly in future smart factories. The following are methods for multiplexing URLLC and eMBB: (1) reserve resources, such as frequency division multiplexing (FDM), and (2) URLLC preempting eMBB resources. However, FDM has low efficiency and preemption impairs eMBB performance. Thus, URLLC preempts eMBB via spatial opportunity, i.e., projecting eMBB to the orthogonal subspace to ensure URLLC performance [85–88].

(3) Complementary TDD. To meet the stringent latency requirement, one solution is to use a self-contained TDD frame structure, e.g., two UL/DL switching points within one slot must meet 4 or 1 ms E2E latency. However, approximately 15% guard period overhead is consumed for a slot format such as [DDDSUUU DDDSUUU], and it is inappropriate to use just one type of frame structure to support services with different latency requirements. Another solution for achieving latency reduction is using an FDD frame structure. However, the DL and UL carriers in the FDD mode are located in two spectrum bands, which might not always be available to a factory operator. Moreover, FDD bands are allocated only with small bandwidth, which might not match the data rate requirements.

The ultimate solution to address the above problem is full-duplex. However, implementing full-duplex at the current stage is extremely challenging. Thus, as a phased implementation step, we proposed

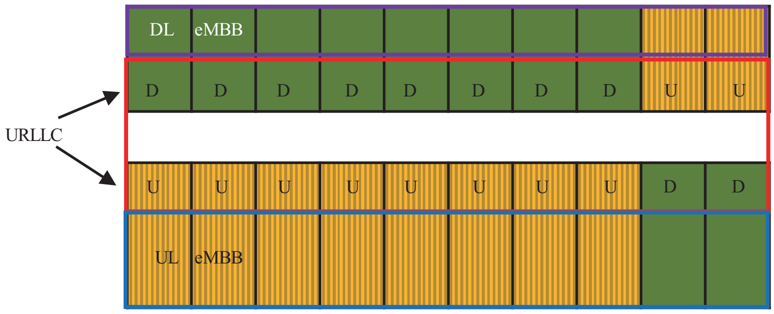

Figure 12 (Color online) Complementary TDD with fully complementary slot formats on non-overlapping frequency sub-bands.

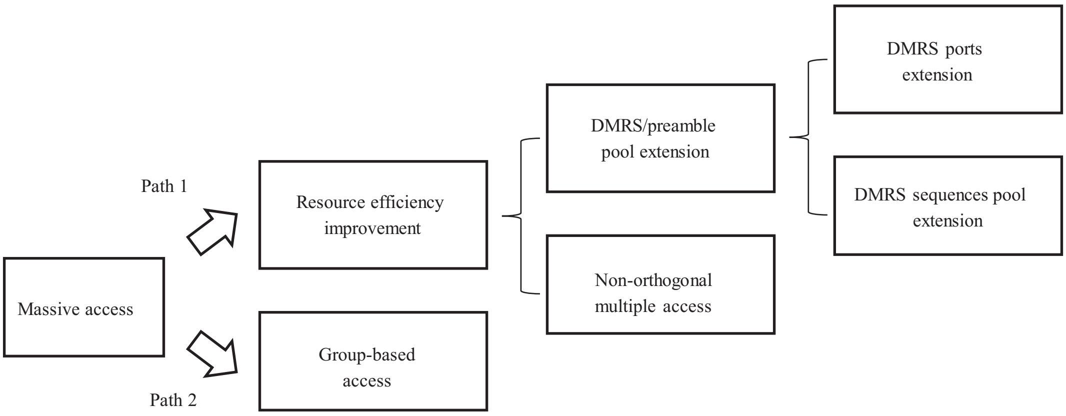

Figure 13 Schemes for massive access.

a scheme called complementary TDD (C-TDD) [89] with fully complementary slot formats on non-overlapping frequency sub-bands to enable simultaneous transmitter (Tx)/receiver (Rx) operation at gNB, which can achieve FDD-like low latency. Figure 12 shows a schematic of the C-TDD. The C-TDD is more feasible than full-duplex, although it still has similar issues to address, such as whether/what filtering methods to be used on the RF or baseband part, whether/how to achieve Tx/Rx antenna separation, whether/how to achieve self-interference cancellation, and whether/how to perform the cross-link interference cancelation.

3.4.3 *Massive access capability*

Before Rel-17 NR specification, the key metrics for mMTC, such as massive access, have not been specifically studied and specified. In 5G-Advanced discussion, the massive access schemes must be explored for the high-density deployment of multiple kinds and a huge number of IoT UEs, such as in ToB scenarios. Massive access can be achieved via these two paths: (1) supporting more UEs using the same time/frequency resources, i.e., improving the resource efficiency, and (2) making fewer UEs represent a large amount of UE access to the network, i.e., group-based access, as illustrated in Figure 13.

For path 1, because of the sporadic and small size characteristic of IoT packages, grant-free and contention-based transmission introduced in Rel-15/Rel-16 is a preferable scheme for massive IoT access. UE can select one DMRS from the pre-configured DMRS pool and transfer the data through the pre-configured UL channels. Here, extending the DMRS pool is feasible for massive access since more UEs can transmit through the same time/frequency resources, and the DMRS contention probability can be decreased with the pool extension. The DMRS pool extension includes ports and DMRS sequence extensions. The DMRS port extension mainly relies on a 3GPP specification design, such as supporting more frequency multiplexed ports, supporting more slots for DMRS, or combining multiple ports to form a new port. However, the DMRS sequence extension can include existing sequence extensions, such as Zadoff-Chu sequences with more roots, new sequence design such as Reed-Muller sequence, fast Fourier transform decimation sequence or AI-based non-orthogonal sequence generation. Our analysis shows that when the number of potential UE is 96, there is approximately a 90% decrease in misdetection probability if the pool size is extended from 12 to 96. Besides DMRS extension, another feasible method to improve

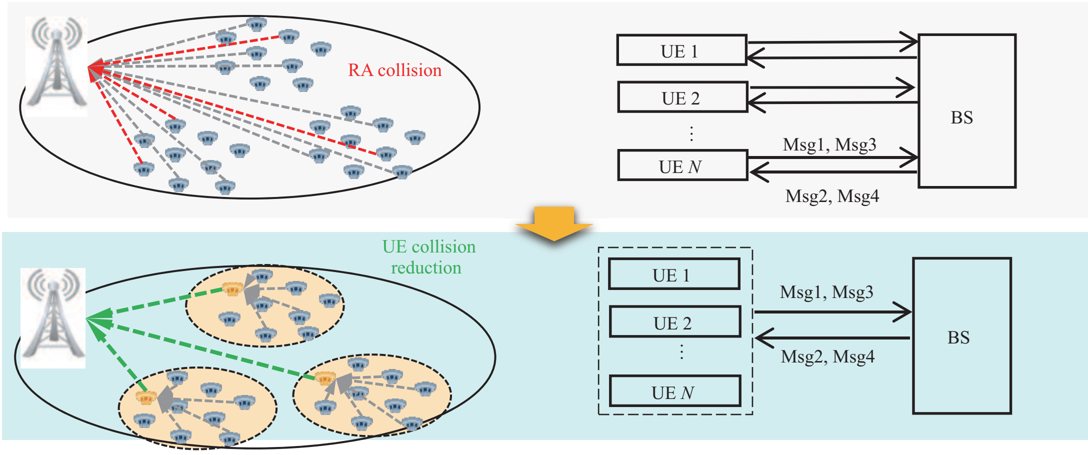

Figure 14 (Color online) Group based access for massive connectivity.

resource efficiency is to support non-orthogonal multiple access, in which more UEs can be superposed in the limited resources within the overloading region.

For path 2, the group-based access based on UE clustering and relaying can reduce the collision and improve the coverage and energy efficiency of IoT UEs, as depicted in Figure 14. The nearby UEs can first be grouped with or without the BS's indication. Then, one or more UEs, such as the group head, can access the network representing several or all UEs in the group. Thus, the collision probability can be reduced. Moreover, the transmit power of group member UE can be reduced since the transmit distance to the group head is significantly shorter than that to the BS. The battery life of UE is not only extended, but the intra- or inter-cell interference can be suppressed. However, the coverage of group member UE can be extended because of the relay through the group head. In the actual deployment, UEs with higher capacity can be chosen as group members, such as smartphones in personal/home scenarios and machines in industries.

Furthermore, the other two enhancement directions for mMTC in 5G-Advanced are further reduced capability and grant-free transmission in the radio resource control (RRC)_IDLE state.

• In 5G NR, on top of the NR UEs defined in 3GPP Rel-15 and Rel-16 for eMBB and URLLC cases, Rel-17 specifies an NR UE type with middle capability/cost between low power wide area and eMBB/URLLC desired by the IoT industry, which is called NR UEs with RedCap [90, 91]. RedCap UEs are designed for UE cost reduction, UE power saving, among others, mainly for MTC application cases such as smart wearable, video surveillance, and industrial wireless sensor network, with data rates of approximately 150 and 50 Mbps for DL and UL, respectively. UE types with further reduced capability would be introduced in 5G-Advanced for the mMTC use cases with peak data rate requirement of DL 2–10 Mbps and UL 1–5 Mbps, such as low-end wearable, shared bicycle, smart agriculture, and industrial sensors. The design purposes of mMTC UEs in 5G-Advanced may not only include cost reduction and power saving but also high accuracy positioning (e.g., sub-meter level), massive connectivity ($>$ 106 UEs within 1 km^2), and coverage enhancement.

• Grant-free, which is a method for UEs to transmit data through pre-configured UL resources, has been introduced since Rel-15 for URLLC services in RRC_CONNECTED, which is called configured grant (CG) in the 3GPP specification. In Rel-17, the grant-free transmissions, such as CG and two-step random access, are specified for supporting the UEs in RRC_ INACTIVE to transmit infrequent and small data packages. In 5G-Advanced, grant-free multiple access can be further extended to RRC_IDLE for further power saving and signaling reduction. The saved signaling resources can be used to access more UEs. Here, the UE's pre-configured resources are shared by the UEs in the cell, and they may be asynchronous with the base station. Technical methods or specification designs for addressing these problems will be studied in 5G-Advanced, such as advanced receiver design for a contention-based and asynchronous grant-free UL transmission.

3.4.4 *Low power capability*

Before Rel-18, 3GPP has specified NB-IoT and NR RedCap to satisfy the requirements on low-cost and low-power devices. These IoT devices consume numerous milliwatts of power during transceiving,

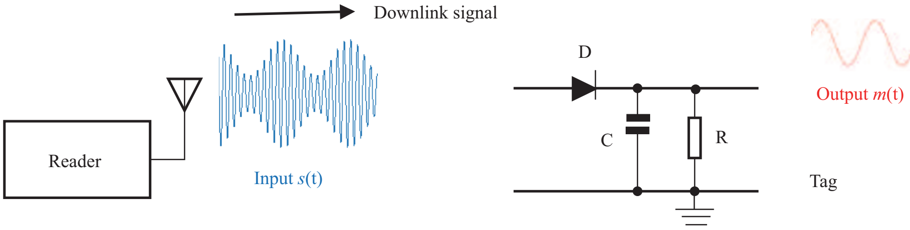

Figure 15 (Color online) An implementation structure of envelope detector.

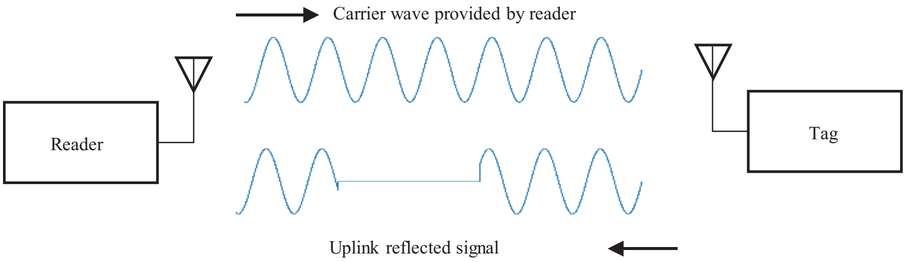

Figure 16 (Color online) Implementation structure of backscatter communication.

but the price of each device is a few dollars. However, to achieve IoE, such as inventory management in logistics, warehousing, manufacturing, and for wireless sensor networks in smart agriculture and industrial automation, devices with ten or even a hundred times lower cost and power consumption is required, particularly for applications requiring batteryless devices.

The key issue for power saving is to avoid using the conventional radio architecture, such as power-hungry RF chains with oscillators, mixers, and digital-to-analog converters. Passive IoT (also known as zero power IoT) is considered a promising technique for satisfying the above-mentioned requirements of device cost and power consumption [92]. To achieve power consumption at the microwatt level, the transceiver of a passive IoT device, such as a tag, can be designed as follows [93].

(1) DL receiver with envelope detection. Non-coherent detection without requiring mixing received RF signal with locally generated carrier waves is the key point for low-power receivers. Among them, envelope detection has the advantage of overall implementation simplicity, with the main part implemented by a diode and a resistor-capacitor oscillator circuit, as shown in Figure 15.

(2) UL transmitter with backscatter communication. In backscatter communication, a reader sends a carrier wave to a backscatter tag as shown in Figure 16. The backscatter tag modulates and reflects the received carrier wave to transmit data instead of generating carrier wave by itself. Communication via reflection rather than active radiation reduces the RF frontend of the tag to a single transistor switch, thus minimizing manufacturing costs and energy demands.

However, there are certain implementation challenges in passive IoT. For example, corresponding signal processing schemes such as coding and modulation must be studied in the above transceiver. Furthermore, hardware capability, such as synchronization, is strongly restricted by the ultra-low power consumption of passive IoT devices. However, the protocol stack must be minimized to simplify the digital circuit and reduce the device size. In summary, a new air interface is required for passive IoT.

3.5 AI-enabled air interface

In recent years, AI and machine learning (ML) have achieved great success in areas such as computer vision, natural language processing, and automotive driving. Applying AI/ML to wireless networks has been being promoted and discussed not only in academia [94] but also in several standardization organizations, such as 3GPP and ITU. In 3GPP Rel-17, there is an active study item [95] exploring the functional framework for RAN intelligence, and various higher-layer use cases for AI/ML have also been identified. In Rel-18, AI/ML will be further studied for NR air interface (i.e., focusing on the physical layer) to enhance performance or reduce complexity/overhead [96].

The potential motivation on AI/ML for air interface is two-folded: handling the design problems that are hard-to-model via traditional approaches and reducing the complexity toward the hard-to-solve problems. To illustrate the advantages of applying AI/ML on the NR air interface, we recognize four

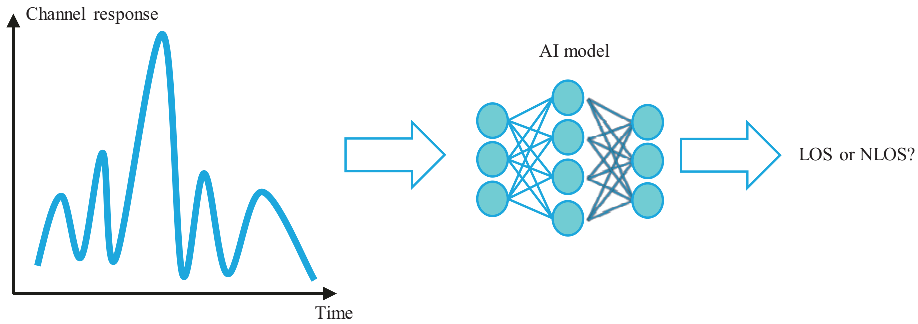

Figure 17 (Color online) AI/ML for LOS/NLOS identification.

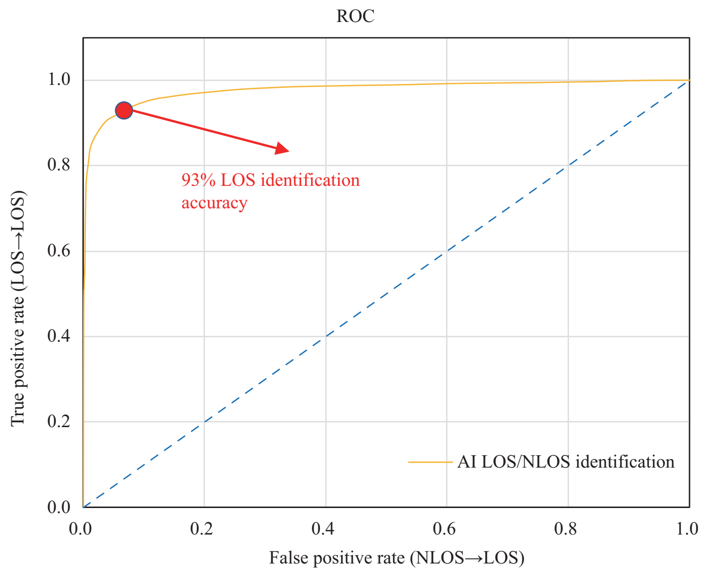

Figure 18 (Color online) AI-assisted LOS/NLOS identification accuracy.

typical and the most discussed use cases that are introduced in the subsequent subsections. Note that the first three use cases, such as positioning accuracy enhancement, beam management, and CSI feedback enhancement, were selected as initial study cases in Rel-18 [96].

3.5.1 *Positioning accuracy enhancement*

In cellular wireless positioning, accurate identification of line-of-sight (LOS) and non-line-of-sight (NLOS) is of great significance for improving positioning accuracy. LOS/NLOS identification corresponds to the binary classification problem. The LOS and the NLOS channels have different channel characteristics, such as multipath delay distribution, energy distribution, and angle spread. Using a single channel feature alone cannot meet classification accuracy requirements because the dominant features suitable for identifying LOS/NLOS are different in various scenarios. Moreover, identifying a proper channel feature is challenging because of the lack of an effective modeling methodology.

Figure 17 shows that AI/ML can be a solution to this problem because of its classification strength by learning the distinct patterns of LOS and NLOS channel responses. Several existing research results show that, based on a large amount of channel data, using ML to train classification models can improve the accuracy of LOS/NLOS recognition. For example, Figure 18 shows that the AI-assisted LOS/NLOS identification scheme can identify LOS channels correctly with a 93% accuracy while misidentifying very few NLOS channels as LOS. Such high recognition accuracy is difficult to achieve using traditional non-AI methods.

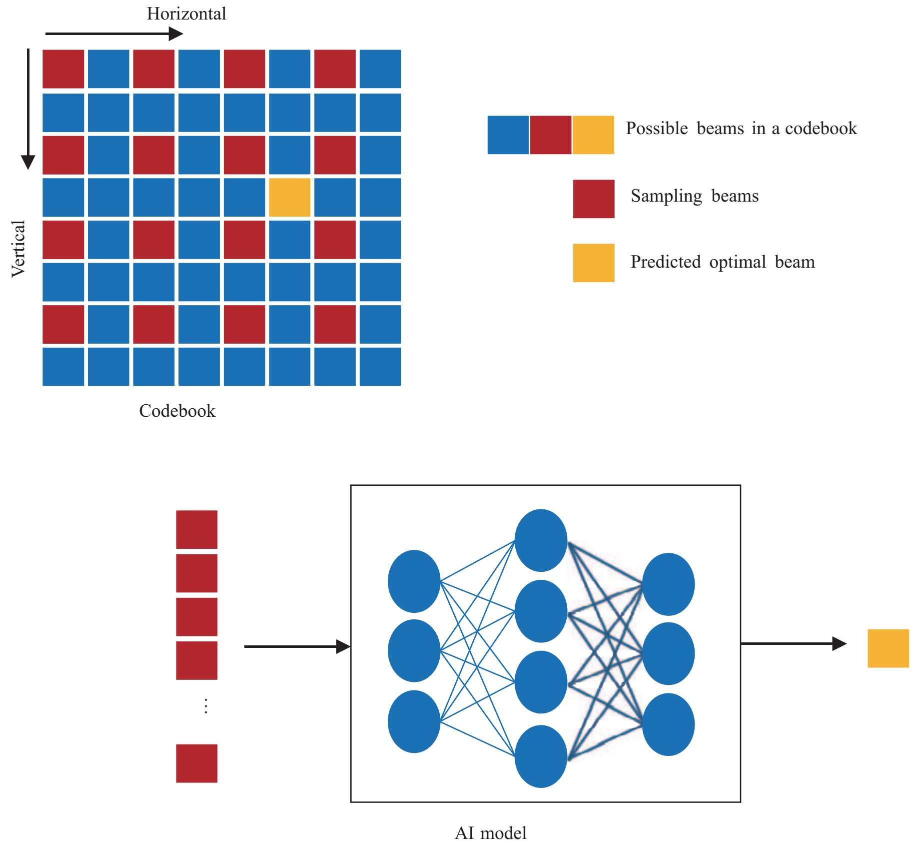

Figure 19 (Color online) Codebook-based sparse spatial sampling.

3.5.2 *Beam management*

In millimeter-wave (mmWave) communications, beam management must align, transmit, and receive analog beams as fast and precisely as possible. Beam management is essential in identifying the main projection direction of the channel by sampling the high-dimensional channel space. Conventional beam-sweeping uses a fixed codebook as the basis for sampling the channel space, and the beam with the largest projection energy among the codebook is chosen for data communication. Beam sweeping treats spatial samplings as independent measurements without using the correlation of successive samplings.

Moreover, because of the channel's randomness, the main projection direction of the channel has energy leakage in each sampling direction, making it possible to predict the main direction of the channel with a reduced number of sampling. Furthermore, AI/ML is applied to perform the prediction because of its powerful feature extraction and prediction capabilities. Figure 19 illustrates this concept, which is known as codebook-based sparse spatial sampling.

Using beam management in the initial access process as an example, where a codebook scanned by the synchronization signal block (SSB) has 64 beams, traditional exhaust-beam-sweeping must scan all beams in the codebook to select the optimal beam. For the proposed codebook-based sparse space sampling, only partial beams (16 beams, for example) in the codebook must be scanned. Through this method, an AI model is trained to predict the optimal beam using the measurements of a part of the beams. The beams used for sparse spatial sampling may be selected in a plurality of patterns, and the pattern can even be studied based on the distribution of UEs, distribution of reflection paths, and channel characteristics. Additionally, beams for sparse spatial sampling may not belong to the codebook used for data transmission, and the wide beam can also be used for spatial sampling.

However, there are several opportunities and challenges in applying AI to beam management. As the number and distribution of channel paths vary significantly in LOS and NLOS scenarios, the codebook-based scheme with fixed beam granularity has limited spatial resolution, which cannot strike a balance between the beamforming gain and the acquisition of channel power [97]. Figure 20 shows that AI/ML can be applied to optimize the analog beams depending on the feedback from UEs, where the UE feedback re-

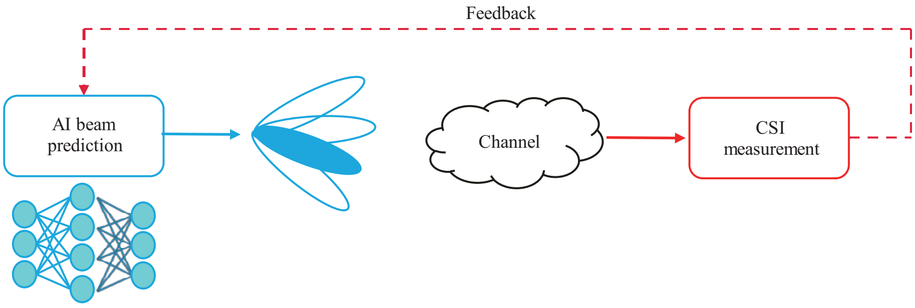

Figure 20 (Color online) AI/ML for beam management.

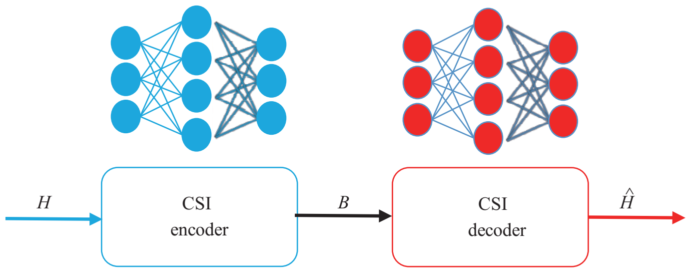

Figure 21 (Color online) AI/ML for CSI feedback.

flects the channel characteristics and the UE mobility. Furthermore, mmWave is an environment-sensitive technology because the changes of multipath would affect beam management significantly. Traditional beam management uses a best-effort strategy, which implies that a beam is selected to match the channel. However, if the channel environment can be modeled in real-time, the blockage and multipath can be predicted accurately; thus, we can perform active beam management, which implies that the beams can be selected as the channel changes. AI's powerful learning capability is suitable for proactive environment sensing. Through perception fusion of multi-band/multi-source data and user trajectory prediction, AI can dynamically model and deduce millimeter-wave environments by implementing delay-free beam steering.

3.5.3 *CSI feedback enhancement*

The acquisition of CSI has always been an eternal topic for achieving MIMO theoretical gains as the accurate acquisition of CSI is crucial for DL data transmission; however, feed-backing the full CSI consumes tremendous overhead in the air interface. Exploiting the sparsity of CSI in the angular-delay domain using mathematical modeling can obtain compressed CSI. However, the optimal converting domain for CSI compression is unknown; thus, the optimal CSI compression is still an open problem.

Recently, AI/ML provided a new solution for high precision CSI feedback compression. Figure 21 shows that AI/ML is applied to obtain the optimal mapping rules between the full CSI and the converting domain, which can be used to build the encoder and decoder for CSI compression and recovery, respectively [98]. As AI is designed to explore the characteristics of the channel space and learn the optimal channel space representation basis, AI-assisted CSI feedback can outperform the model-based CSI feedback in Rel-17. Through the joint training of the UE encoder and the BS decoder, AI-assisted CSI feedback can achieve 110% higher throughput compared with Rel-17.

AI-assisted CSI compression feedback has other potential research directions and challenges. Current research mainly focuses on CSI feedback of a single UE. There is a correlation between the channels of multiple UEs. CSI feedback overheads can be significantly reduced if channel compression can be performed together with multiple UEs. Additionally, the CSI feedback may further be combined with channel prediction, and joint CSI of a plurality of timeslot or symbols is fed back by leveraging the correlation between channels in the time domain. Moreover, multi-band joint CSI extrapolation and prediction is also a promising research direction. Considering that a UE can communicate on multiple frequency bands, joint CSI compression and feedback on multi-bands may facilitate optimal UE scheduling and interference management.

Despite the aforementioned promising application cases, AI/ML on air interface also suffers from

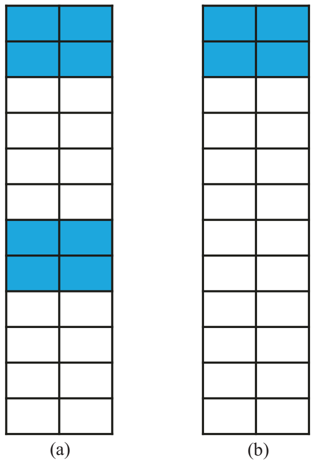

Figure 22 (Color online) DMRS configuration. (a) 4RE/RB; (b) 2RE/RB.

several challenges. Because the training samples of wireless networks are noisy, diverse, and difficult to collect in different scenarios, guaranteeing the robustness and generalization of AI/ML models is challenging but highly desired. In 5G-Advanced, this issue might be addressed by carefully designed signaling and advanced AI techniques, such as transfer learning, semi-supervised learning, and mate-learning techniques. However, the employment and adaptation of advanced AI techniques on wireless problems require further study [99, 100].

3.5.4 *Channel estimation and RS design*

Based on the current 5G NR design, improving the UL capacity is challenging because orthogonal pilots limit the number of data streams that can be supported, and the pilot overhead reduces the resources that can be used to transmit data. Based on these problems, AI provides two approaches to overcome these limitations: AI channel estimation for sparse DMRS and joint RS design and channel estimation.

(1) Sparse DMRS and AI-assisted channel estimation. Figure 22(a) shows a typical DMRS configuration supported by 5G NR: the DMRS of each data stream occupies two symbols in the time domain and four REs for each RB in the frequency domain. Figure 22(b) shows a sparser DMRS configuration; compared with Figure 22(a), only two REs are occupied in each RB, thus reducing DMRS overheads or supporting more data streams. To achieve the desired performance for sparse DMRS configuration, two AI-assisted channel estimation schemes are considered as follows.

Scheme 1. A channel estimation algorithm based on a convolutional neural network (CNN). Similar to [101], Scheme 1 employs CNN to perform high-resolution recovery and denoising on the roughly estimated channel. First, the least square (LS) algorithm is used to estimate the channel of the RE where the DMRS is located. Then, the roughly estimated channel is input into a CNN-based multilayer residual network to obtain the denoised channels of all REs.

Scheme 2. A channel estimation algorithm based on the learned approximate message passing (LAMP). Similar to [102], Scheme 2 adopts the method for combining the AMP algorithm and the neural network, expands each iteration of the algorithm into a layer of the neural network, and learns the algorithm's parameters through model training. In this method, the LAMP algorithm is used to estimate the channel in the delay domain according to the DMRS and the received signal, and then the channel of all REs in the frequency domain is obtained using DFT.

Common channel estimation algorithms include the LS and linear minimum mean square error (LMMSE) algorithms. This study shows that the LS algorithm has the worst performance because the noise and the channel statistical characteristics are not considered. The ideal LMMSE algorithm takes advantage of the statistical characteristics of the noise and channel; however, these characteristics are extremely difficult to obtain in practice. The AI-based channel estimation algorithm can learn these statistical characteristics during the training phase. Compared with the LS algorithm, the performance of the two AI-assisted algorithms is significantly improved, and the performance is robust for two different DMRS configurations.

(2) AI-assisted RS design and channel estimation. When there are massive UEs for UL grant-

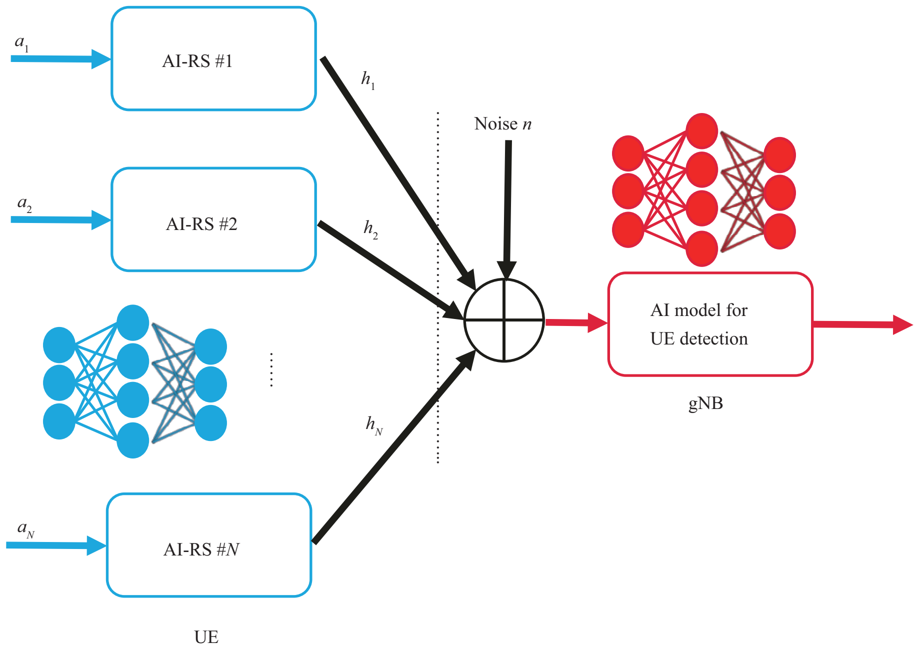

Figure 23 (Color online) AI/ML for RS design and channel estimation.

free transmission in the same resource, which is significantly more than the number of orthogonal DMRS ports, the RS design problem and UE detection become quite complicated. Figure 23 shows that AI/ML can be applied to obtain the RS pool and corresponding UE detection algorithm efficiently without searching and testing all candidate RS patterns and sequences [103]. Such a design can significantly increase the potential UL capacity, and it requires further study to fully unveil the gains.

4 Technological directions in spectrum domain

4.1 Harmonized communication and sensing

HCS, also known as integrated sensing and communication, combines sensing with communication systems and pursues mutual benefits to use wireless resources efficiently in applications such as vehicular networks and smart cities. Note that HCS has been widely recognized as a 6G enabling technology. Because of space limitations, we only discuss one root technology in the HCS domain, such as location technology, and will use V2X as an example to explain wireless sensing.

4.1.1 *Low power high accuracy positioning*

5G positioning technologies [104, 105] provide better support for several industries in terms of enterprise management, security monitoring, emergency rescue, trip monitoring, among others. These use cases require not only high accuracy positioning but also low power positioning; thus, low power high accuracy positioning (LPHAP) has already become the next evolution direction of positioning technology. The key to reducing power consumption is minimizing the communication capability. For example, in the case of asset tracking, the service that the device must provide is its location with various horizontal accuracy targets; however, communication capability can be minimized to support positioning-related functions, such as UE initial access, network/security authentication, synchronization, and battery volume reporting, if required.

Among various positioning methods, UL time difference of arrival (UL-TDOA) is the most effective for achieving LPHAP because the UE operation only synchronizes the network and send/receive RS[2] and otherwise goes into a deep sleep to save power. The following are four potential techniques identified for reducing device power consumption while maintaining high accuracy.

2) In the NR specifications, these signals are the positioning reference signal in the DL and SRS for positioning in the UL.

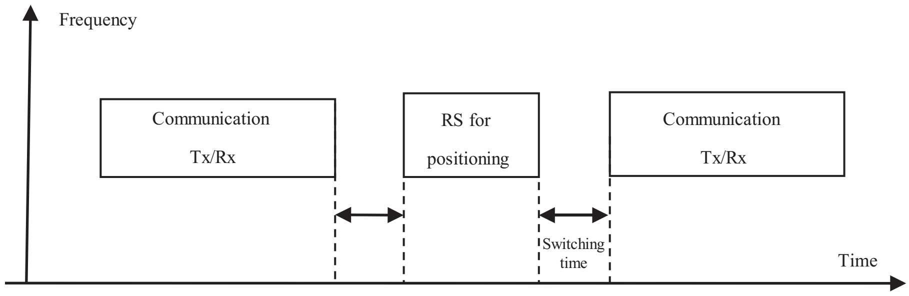

Figure 24 LPHAP bandwidth adaptation in UE side.

(1) Decoupled communication and positioning bandwidth. LPHAP is expected to support narrowband communication and wideband positioning simultaneously, thus allowing a device to function in a narrowband mode with low power communication while operating in a wideband mode for positioning with high accuracy. Figure 24 shows that switching time should be reserved to allow for device switching between the two working modes.

(2) Mobility enhancement. Rel-17 is currently working on RRC_INACTIVE positioning with UL and UL plus DL positioning. However, supporting positioning in the RRC_INACTIVE state is insufficient for achieving a long battery life of one or more years. However, what matters is the mobility problem, which should be further optimized. For instance, when UE accesses a new gNB with connection resumption due to mobility, the network releases the SRS allocated in the source gNB and configures a new SRS configuration to the UE so that it can continue transmitting SRS. However, the time delay of UE re-entering RRC_CONNECTED status and receiving a new SRS configuration could be lengthy since the target gNB, source gNB, access and mobility management functions, and location management function must exchange the UE context for updating the SRS configuration, resulting in less sleep time and high power consumption for UEs. Defining a positioning area is a method for avoiding the re-configuration procedure, which can help save UE power. When moving across multiple cells within the same positioning area, UE can maintain the SRS configuration for positioning.

(3) Ultra-deep sleep. Keeping UE in the sleep state for a long time and reducing the power consumption of the sleep state are the most straightforward ideas for UE power-saving. UE power consumption evaluation in 3GPP is based on deep sleep with the assumption of one power unit per slot, 20 ms transition time, and 450 power units during the transition time. However, such a deep sleep state is insufficient to support the long battery life required by LPHAP, and even lower power consumption in an ultra-deep sleep state should be introduced.

(4) Aggregated SRS. The accuracy of UL-TDOA crucially depends on the bandwidth of SRS, which is limited to 100 MHz in sub-6 GHz. Aggregated SRS is an efficient method for extending the bandwidth. According to Cramer-Rao bound analysis, the error of time of arrival estimation of the first arrival path is proportional to the equivalent bandwidth (the bandwidth between the lowest and highest frequencies in the aggregated SRS). This allows the joint processing of SRS in contiguous or non-contiguous bands to achieve higher accuracy. For example, aggregated SRS in two narrow bands, each with 20 MHz and a 60 MHz gap in between, can achieve peer accuracy of 100 MHz contiguous band if the phase error between the two narrow bands can be estimated and ideally compensated. UL-TDOA with the non-contiguous bands is promising to achieve sub-meter level accuracy.

Another aspect to note is the unideal factors in practical engineering, such as inter-site synchronization error and inaccurate network parameters that significantly affect the positioning accuracy, thus wasting the UE energy. More robust positioning techniques such as that based on the virtue of UE/BS reference with known location should also be considered in LPHAP.

4.1.2 *Wireless sensing for V2X communication*

In addition to high-accuracy positioning, the outstanding sensing capabilities offered by various cutting-edge sensors (e.g., mmWave radar, light detection and ranging, and red, green, and blue (RGB) camera) can also benefit V2X communication. We envision that today's vehicles are no longer stand-alone transport because advancements in vehicle-to-vehicle and vehicle-to-infrastructure communications have transformed them into intelligent nodes for information interaction in wireless networks. Thus, wide-ranging vertical applications such as real-time HD map delivery and distributed sensor sharing based on

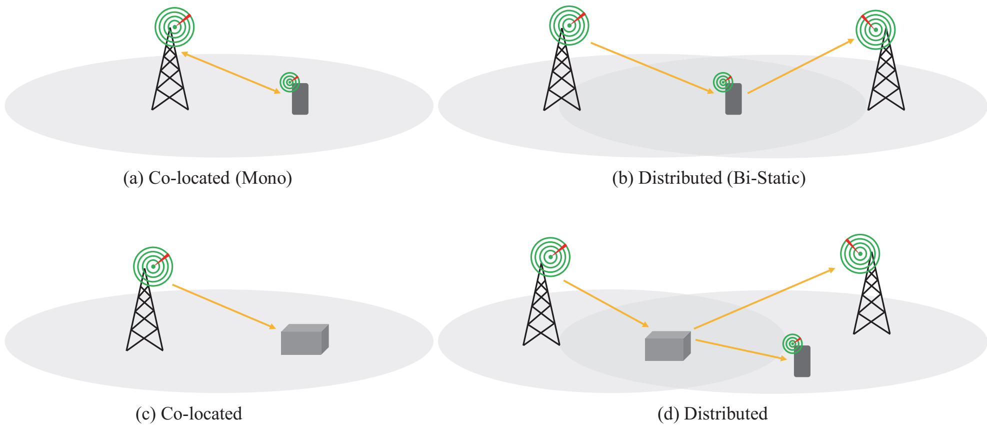

Figure 25 (Color online) RF based sensing: (a) and (b) device-based sensing; (c) and (d) device-free sensing.

advanced intelligent transportation systems (ITS) can be implemented to improve driving safety and the operating efficiency of the vehicle network.

HCS can be achieved in different ways. Numerous sensing techniques have been developed in the last decades, such as radio imaging-sensing. This study focuses on RF-based sensing, as shown in Figure 25, which uses communication signals for sensing purposes and turns cellular network into a networked sensor. RF-based sensing can be classified into two main categories: device-based and device-free sensing. The former targets registered network elements such as mobile phones and vehicles, whereas the latter deals with unregistered objects such as buildings and trees, which is similar to the traditional radar system [106]. Furthermore, sensing functions can also be categorized based on their system architectures. If the transmitter and receiver are co-located, it is called monostatic sensing; otherwise, the sensing is considered bi-static if the transmitter and receiver are separately located in space. Additionally, the integration of sensing and communication functions can be divided into several levels. The lowest level only shares spectrum, and if we consider a little further, the sharing can involve hardware, RF, digital signal processing, PHY processing, and protocol interface. Through these methods, flexible information exchange among different layers, modules, and nodes can be achieved in V2X communications. Ultimately, the highest level enables tighter integration so that we can design a single system that mutually enhances the performance of sensing and communication. Note that HCS does not simply mean to use communication signals for sensing, but the sensing functionality can also help wireless communication, such as cell planning, beamforming, and channel acquisition. HCS opens a new approach to use the information from these two separate functions and significantly enhances the current vehicle networks from every perspective.

With the help of wireless sensing, high-rate data exchange and high-precision localization can be achieved in V2X communication. Specifically, sensing information can be used to perform high-accuracy beam tracking and prediction in mmWave V2X communication, which can reduce the training overhead and improve spectrum efficiency. Furthermore, wireless sensing can also gather environmental information, which is particularly useful for applications such as vehicle platooning, forward collision warning, and intersection movement assistance. These advancements will eventually accelerate the development of L4/L5 autonomous driving, which will benefit the entire automotive industry.

Despite remarkable advantages, the current HCS-assisted V2X communication is still challenged by several fundamental issues that must be adequately addressed before commercialization. First, realizable KPIs with cost and power consumption constraints must be properly defined. In wireless communication, KPIs such as bit and block error rates are defined based on mathematical models, and they have been demonstrated in various tasks. However, new KPIs, such as sensing accuracy and resolution, have not been adequately defined in HCS-assisted V2X communications. Second, the fundamental limitations and trade-offs of HCS are still unknown. It is crucial to provide a comprehensive performance analysis to evaluate the superiority and guide the design of the HCS-assisted V2X systems. Furthermore, the performance-complexity trade-off must be considered since it is a classic problem in every scenario. Third, new channel modeling and evaluation methodology must be developed. Stochastic channel modeling is widely adopted in the design of wireless communication systems since it supports fast software simulations. However, such a mechanism may not be applicable, especially in localization, positioning, and mapping applications, since their performance depends heavily on the wireless environment. Finally, the system-level design of HCS-assisted V2X communication must be considered. Theoretically, multiple HCS-assisted vehicles can be grouped functionally to achieve a joint perception of a target in a given area. In this method, multiple sensing tasks such as high-resolution imaging and positioning can be fulfilled. Here, it is necessary to study the information exchange and cooperative sensing processes between vehicles [107].

While research into HCS-assisted V2X and other aspects of vehicle interaction are still in their early stages, it is crucial to establish a clear vision for future V2X technology to guide academic researchers and industrial partners. Recently, the IEEE Communication Society announced an emerging technology initiative (ETI) to explore and support various research directions and standardization opportunities related to HCS [108]. We assume that this ETI will provide a great opportunity for researchers in academia and industry to openly present new ideas and exchange technical insight on HCS-assisted V2X communication.

4.2 High frequency (HF) capability enhancement

The 5G spectrum does not only cover bands below 6 GHz but also extends into much higher frequency bands of approximately 100 GHz (also known as mmWave). The HF bands can provide large bandwidths and support high data rates that are ideal for increasing the capacity and reducing the latency of wireless networks. To cope with the challenges of HF deployment primarily caused by high penetration losses, blocking, and high cost, HF technological evolution will be an important part of 5G-Advanced. In our vision, a large array, flexible duplex, and multi-frequency coordination will be the most promising technological directions for HF capability enhancement.

4.2.1 *Large array*

Although HF support has been introduced since Rel-15, compared with the widely popularized deployment of 5G NR over low frequency (LF), commercial deployment of HF today is still limited because of low coverage and high cost. Using a large array for analog beamforming at gNB and/or UE is a possibility, while the number of digital transceiver units remains unchanged [109].

In terms of deployment cost, it is more efficient to deploy HF reusing existing site locations obtained for LF, which requires that macro HF gNB must be capable of providing continuous coverage for a typical inter-site distance (ISD) from 300 to 400 m. To this end, our preliminary field trial shows that a 1024–4096 antenna array is required at gNB to guarantee a targeted DL data rate of 1 Gbps. As for UE, because of space limitations on handheld devices, one antenna panel at UE is assumed to have approximately eight elements. These limitations may be less applicable to non-handheld devices, such as customized devices for live broadcasting or real-time HD video uploading. For these types of devices, a larger array can be considered, such as 16 to 32 elements per panel. These large arrays at UE are particularly suitable for UL-dominant scenarios. Our preliminary field trial (also under 300 to 400 m ISD) shows that a 16–32 antenna array is required at UE to guarantee a targeted UL data rate of 30 Mbps.

With obvious benefits from using large arrays at HF, the following challenges should be addressed.

• Mobility with narrow beams due to a large array. Figure 26 shows that using a large array at gNB for data transmission toward remote locations increases more frequent beam switching. Additionally, the narrow beams result in a sharper fall-off deviating from the peak direction, which is more sensitive to UE movement and rotation. Similarly, narrow UE beamforming resulting from a large array at UE will also create more challenges in mobility support, such as moving vehicles.

• Furthermore, the common problem of blockage and UE rotation is also present and more serious because of narrow beams. The uncertainty of beam-based communication is mainly caused by a blockage. When a blockage occurs, the channel quality drastically deteriorates, and beam failure occurs. The rotation of UE (caused by changing of grasping posture) results in unpredictable beam switching, which requires fast switching of UE panels and potentially re-selection of gNB serving beams.

• Energy saving at gNB and UE. A narrower analog beam means that gNB will transmit data in the time-division multiplexing method, and the energy consumption would unavoidably increase if not handled properly. Furthermore, UE must perform more frequent beam measurements and continuous maintenance; thus, the energy consumption at HF UE is expected to be even more challenging.

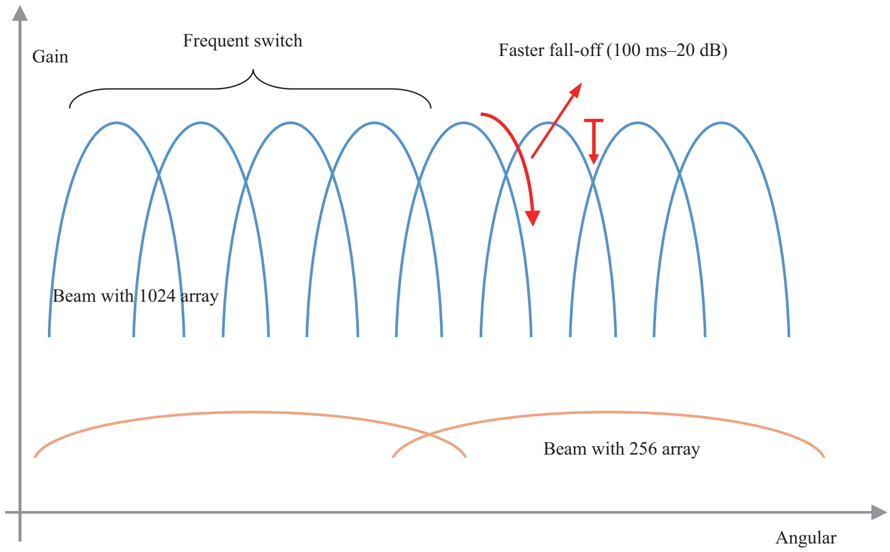

Figure 26 (Color online) An illustration of challenges in beam-based mobility with large-array HF gNB.

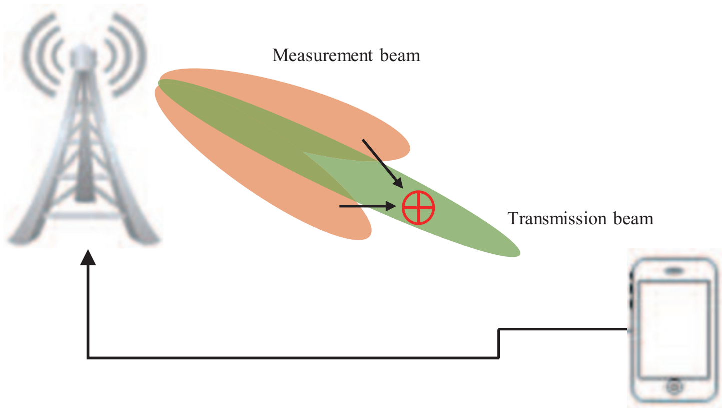

Figure 27 (Color online) UE-assisted gNB narrow beamforming with wide measurement beams.

To support a large array, particularly to improve mobility support under narrow beamforming, the following directions can be investigated[3)].

• UE-assisted gNB narrow beamforming. Although gNB may form a narrower beam for data transmission, it does not necessarily mean that the candidate beams used for beam measurement would also be narrower. Thus, we can introduce additional reporting, such as phase difference among multiple reported gNB Tx beams to enable gNB to form a narrower beam with reduced or even without sweeping of narrow beams, as shown in Figure 27.

• gNB-assisted UE beam refinement. In the current specification, gNB does not have significant control over UL beam training at the UE side. For example, gNB can only instruct UE to use a Tx beam that is the same as a previous Rx or Tx beam; however, it cannot permit UE to use a wide Rx beam (DL SINR saturated) but a narrower Tx beam (for higher gain). In such cases, it would be beneficial to allow the gNB to task the UE to further narrow its Tx beam relative to the Rx beam, as shown in Figure 28.

4.2.2 *Flexible duplex*

In the current 5G NR deployment, most of the spectrum is unpaired using TDD duplex mode because large bandwidth is introduced. Moreover, a flexible duplex such as a flexible UL/DL slot ratio can be configured for a cell in a TDD carrier to better support diverse traffic requirements. However, a flexible

3) We will discuss energy-saving solutions in Subsection 5.1.

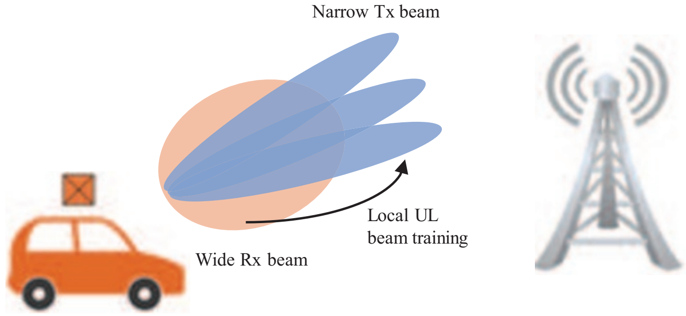

Figure 28 (Color online) gNB-assisted UE beam refinement with asymmetric Rx and Tx beams.

duplex is still far from real implementation because of the following two main bottlenecks that require further academic and standardization studies in 5G-Advanced.

(1) Cross-link interference mitigation. In a flexible duplex, neighboring BSs may use different transmission directions in the same spectrum and time, where cross-link interference (CLI) is a severe problem that must be resolved. In UL, the base station receives interference from neighboring BS DL transmission, which is referred to as DL-to-UL or BS-to-BS interference. In DL, UE receives interference from neighboring UE uplink transmission, which is referred to as UL-to-DL or UE-to-UE interference. Various approaches have been proposed to address the CLI issue. These schemes are classified into coordination-based and receiver-enhancement schemes. For the former, multiple BSs or UEs cooperate to avoid or mitigate CLI such as clustering [110,111], coordinated scheduling [112,113], coordinated power control [114,115], coordinated beamforming [116,117], and UL/DL configuration [118,119]. For the latter, CLI is mitigated by advanced interference suppression and cancelation algorithms at a receiver, such as interference rejection combining via spatial dimension [120, 121], joint maximum likelihood detecting signal and interference [122], and interference cancelation [123].

(2) Out-of-band leakage mitigation. In a flexible duplex, eliminating UL and DL interference within a carrier is a crucial technical challenge because the non-ideal PA causes severe out-of-band leakage. The primary solutions to these issues include enhanced filters [124] and adjacent channel interference cancelation [125, 126]. The former provides better out-of-band isolation, whereas the latter reconstructs or extracts the out-of-band emissions and then subtracts them from the received signals.

4.2.3 *Multi-frequency coordination*

Communication over LF and HF bands has different characteristics. LF band has carriers with small bandwidth, i.e., from several MHz to 100 MHz, whereas the HF band has carriers with larger bandwidth, i.e., from 100 MHz to several GHz, which can support extremely high data rate transmission. However, HF communication also poses various challenges because of high propagation loss, weak scattering and diffraction, and high RF complexity. Furthermore, the transmission over HF is more sensitive to blockage. Additionally, large beam sweeping overhead causes high UE power consumption, and low PA efficiency results in poor coverage, especially for UL. Thus, multi-frequency coordination will be a promising direction for increasing system spectral efficiency in 5G-Advanced by jointly taking advantage of different characteristics of multiple frequency bands.

Multi-frequency coordination techniques in the following areas could be considered in 5G-Advanced to reduce UE power consumption, improve user experience data rate, and decrease system overhead.

(1) Smart offloading between multiple frequency bands. Because the LF band has a wider coverage and higher channel reliability than the HF band, it is crucial to understand how to efficiently and dynamically offload traffic of different UEs with different data rates, latency, and reliability requirements on multiple frequency bands to obtain good user experience data rate. For instance, intra-frame and predictive-frame (P-frame) of XR video streaming could be scheduled on LF and HF band, respectively; hence, it will be less harmful if the P-frame transmission, which is less important, fails because of sudden blockage.

(2) Using correlations between multiple frequency bands. With advanced RF devices and accurate calibrations, power/time/spatial correlations between LF and HF bands could be used in a collocated deployment. It will be beneficial to consider LF carriers as anchor carriers from which HF carriers can obtain large-scale channel states, such as the reference signal received power (RSRP), and

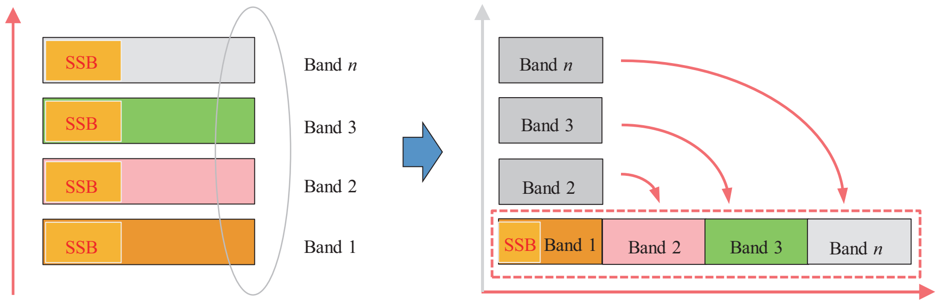

Figure 29 (Color online) MB-SC reconstructs multiple carrier bands into one serving band logically.

small-scale channel states, such as coarse power delay profile (PDP)/power angular spectrum (PAS) of the best beam, which leads to reduced UE power consumption and decreased system overhead.

• Power. For multiple frequency bands that are collocated, the base station could determine the addition/release of HF cells by roughly estimating the RSRP of the HF band with that of the LF measurement band using the Friis transmission equation, which avoids radio resource management (RRM) measurement for the HF band.

• Time. Correlation between LF and HF bands in the time domain was analyzed [127], and it was discovered that blockage caused by appearing obstacles can be observed in the LF band earlier than in the HF band. By this means, users can search in advance for candidate beams that correlate lowly with the current beam, thus preventing users from beam failure and guaranteeing an uninterrupted good user experience.

• Spatial. Spatial correlation between LF and HF bands were explored in [128–130], showing that PDP/PAS/covariance matrix is correlated with different frequency bands using theoretical analysis and measured data. Additionally, our measured results show that if the LF and HF antennas are non-coplanar with common normal, angle profiles have a high correlation of approximately 70%–90% and 30% for LOS and NLOS, respectively. If the LF and HF antennas are coplanar with a common normal, the correlation of angle profiles for different frequency bands may be higher. Based on the correlation, CSI of the HF band can be extrapolated by those in the LF band, which can significantly reduce beam sweeping overhead and improve the CSI accuracy of the HF band.

4.3 Spectrum value maximization

To maximize the spectrum value, 5G-Advanced evolves to use the sub-100-GHz full spectrum through full-band intelligent aggregation and UL/DL decoupling.

4.3.1 *Multi-band serving cell*

The FDD spectrum is a non-contiguous large aggregated bandwidth. For example, in Europe, a bandwidth of 95% contiguous FDD spectrum is $\leqslant$ 30 MHz, mainly consisting of 5, 10, and 20 MHz. Additionally, 90% of operators own more than one FDD band. To fully use the FDD spectrum, the current mechanism uses the CA mechanism. The following two performance characteristics are affected. (1) Each scattered FDD carrier must send common control channels such as broadcast and paging, which leads to excessive common channel overheads and increases the energy consumption of the base station. (2) In CA mode, processing techniques, such as measurement, synchronization, and activation per carrier are required, takes ten milliseconds, and affect user experience. To solve this problem, a multi-band serving cell (MB-SC) is proposed to reconstruct multiple FDD scattered carriers into a virtual large carrier. In this mode, bandwidth part (BWP)[4] is used to implement flexible scheduling among FDD carriers, reducing common overheads, and improving user experience, as shown in Figure 29.

In MB-SC, for UEs in IDLE/INACTIVATE state, system information (e.g., SSB and system information block (SIB)) and paging of a serving cell can be broadcast only in one of the multiple DL carriers within the serving cell. The system information includes necessary information (e.g., frequency, bandwidth, and subcarrier spacing) of each carrier. Thus, all carriers are visible and reachable by UEs. Based on the SSB and SIB, UE can choose any of UL carriers for initial access without secondary cell (SCell)

4) BWP is a contiguous set of PRBs on a given carrier introduced in 3GPP Rel-15 for dynamically adapting the carrier bandwidth and numerology in which a UE operates.

addition and activation procedures. Through this method, MB-SC can reduce the overhead of common signaling, achieve load balancing on different UL carriers, reduce the access delay, and improve the network flexibility and energy efficiency.

UEs in a CONNECTED state can use the frequency resources from multiple bands in the way as a serving cell. The network can control UE's RF and baseband (BB) within a carrier to save power when there is no traffic transmission. This carrier can provide basic signaling for UE synchronization and channel measurement. Multiple BWPs at multiple bands can be configured to the UE. Because of the neighboring bands in a cell that shares the same synchronization and have similar channel states, 1–2 ms BWP-based operation (e.g., BWP activation and switch) can be used to quickly adapt UE's RF and BB for aggregating multiple carriers to achieve high data rate when burst traffic arrives. Activating/deactivating the SCells is unnecessary, i.e., there is no procedure or latency for SCell search, time and frequency synchronization, and RSRP measurement for automatic gain control (AGC) settling. This method reduces the latency, achieves instant wideband transmission, and improves user-perceived data rate.

To achieve the above benefits, MB-SC must overcome the following technical challenges at the system level.

- Broad-band transceiving across multi-carrier involves two technical challenges: (1) improving the wide-band efficiency of RF components, which involves research on new materials and processes, and (2) eliminating cross-modulation interference.
- Accurate time-frequency synchronization between carriers is the basis for flexible scheduling between FDD carriers. The clock sharing architecture and time-frequency offset compensation algorithm must be further studied to improve the precision of time-frequency synchronization between carriers.
- The large-scale information between FDD discrete spectrums is consistent; thus, research on multi-carrier channel information mutual aid techniques, such as multi-frequency channel acquisition and signal processing based on compressed sensing or AI, can improve the performance of MB-SC.

4.3.2 *Flexible spectrum access*

To address the shortness of TDD spectrum coverage in the UL, 5G proposes an innovative technology of UL and DL decoupling. However, the 5G decoupling mechanism only works on a single frequency band, which cannot be applied directly to multi-band scenarios. Moreover, the addition of more frequency bands and the use of current technologies will inevitably increase the terminal transmission capability.

To address the preceding problems, we propose a technical solution of flexible spectrum access (FSA). On the premise of not increasing the complexity and cost of terminal implementation, flexible switching of transceivers (TRXs) enables flexible access of terminals on all frequency bands. This increases the experience rate and reduces power consumption. Specifically, the capabilities of configuration, activation, and simultaneous transmission of a UE are decoupled. Using UL as an example, shown in Figure 30, a UE with only two concurrent Tx RF chains can be configured, activated, and scheduled on more than two frequency bands while performing the concurrent data transmission on only one or two of those bands. Therefore, FSA provides a mechanism for dynamically selecting a subset of configured carriers and accordingly switching Tx for transmission based on the traffic, TDD configuration, and channel condition of each band rather than switching using RRC-based cell(s) or cell group(s) reconfiguration, which requires much longer delay to perform such a selection/switch. Thus, the benefits of FSA are as follows.

- Higher UL capacity. The network can dynamically schedule the UL on the bands within wider bandwidth with the most unscheduled RBs or better channel conditions in a given slot. For example, when one of the active TDD bands/cells is DL for a given slot, the UE can be switched to another TDD band that is UL based on the TDD configurations, an FDD band with the best channel condition, or a band/cell with most spare RBs that can provide higher UL data rate; the UE can be switched back to the TDD band when UL slot is available. Thus, FSA can achieve a higher UL data rate because of more UL available slots, frequency resources, and better frequency diversity.
- Lower UL latency. Emerging latency-bounded traffic such as XR imposes strict latency (millisecond-level) and reliability requirements. FSA can alleviate the transmission timeout issue and offer a considerable performance improvement for these applications because of additional UL available slots. For example, UE can be switched to the UL band in the UL slot if the current band is in a DL slot or congested; however, the UE can be switched back if the original band switches to a UL slot or has more unscheduled RBs.

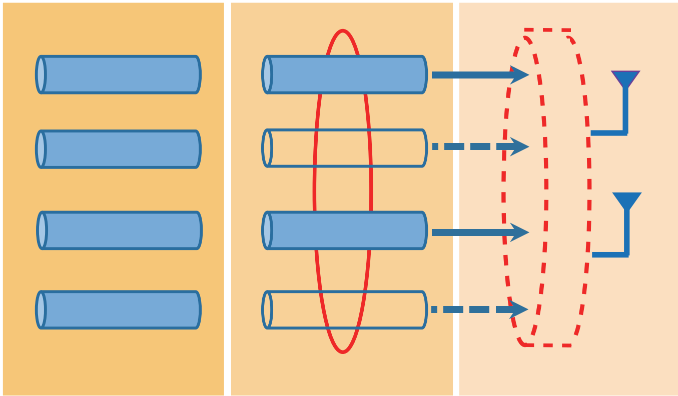

Figure 30 (Color online) Conceptual illustration of FSA.

• Higher spectrum usage. FSA can achieve a well-balanced system load by enabling fast carrier switching on a transmission time interval level. Considering that the traffic arrives randomly if a frequency band is congested with user traffic, FSA can dynamically allocate a part of the traffic load to another frequency band to use the unoccupied resources as much as possible. Thus, higher system spectrum usage can be achieved.

However, flexible spectrum switching also increases system complexity, and additional system overheads such as UL and DL channel estimation and user resource scheduling are introduced. AI may be a feasible solution to address these problems by learning from wireless communication systems. Specifically, AI can effectively reduce the channel estimation overhead on the premise that the radio channel frequency bands are close, e.g., interpolation and extrapolation using channel correlation [131–133]. Moreover, AI algorithms [134–136] can be used to learn user characteristics and achieve resource allocation with low complexity, low latency, and high reliability in a complex network environment with the future interconnection of everything, thus maximizing time-frequency domain resource use.

5 Technological directions in network domain

5.1 Network energy efficiency optimization

According to a report from the global system for mobile communications (GSM) association [137], the energy consumption of information and communication technology (ICT) has significantly increased to approximately 2%–3% of the global energy consumption. This large energy consumption has incurred a large amount of carbon emission; hence, the ITU has proposed a goal to reduce the ICT emission by 45% from 2020 to 2030 [138]. Furthermore, GSMA reports that the energy cost on mobile networks accounts for 23% of the total operator cost, and most of the energy consumption comes from RAN. Using the China Mobile Communications Corporation as an example, the bill for energy reaches approximately 3.8 billion CNY in 2020 [139]. Hence, it is significant to study the optimization of network energy efficiency to reduce network energy consumption for both environmental sustainability and operation cost savings. For many years, 3GPP has been developing energy efficiency standards for mobile networks and proposing several network energy-saving technologies such as carrier shutdown, channel shutdown, symbol shutdown, deep sleep, and symbol aggregation shutdown [140]. Although the power consumption per unit of traffic is extremely reduced in 5G, 5G power consumption is significantly higher than that of 4G [141]. Thus, it is crucial to further optimize the network energy efficiency and reduce the total energy consumption for the 5G networks.

5.1.1 *Energy consumption distribution of network*

The optimization of network energy efficiency minimizes the total network energy consumption per bit transmission. However, such optimization does not consider the network/user performance, and the optimal energy efficiency is achieved at the cost of significant performance loss, which is unacceptable

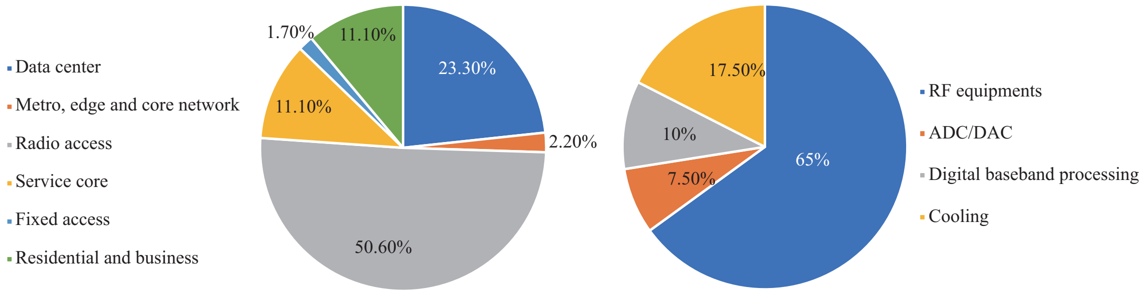

Figure 31 (Color online) Energy consumption breakdown by network element (left) and by the modules within radio access (right).

for the operators. To this end, the study should focus on how to use/efficiently allocate the energy to achieve the best trade-off between the total energy consumption and the network/user performance loss. To achieve this, the first step is to understand how energy is consumed in the current network.

Figure 31 shows the distribution of the current network energy consumption. The energy consumption of the current mobile network mainly comes from radio access such as base stations that occupy approximately 51% of the total energy consumption. Additionally, the RF equipment consumes approximately 65% of the total energy within a base station [142]. Currently, the C-RAN architecture is widely used, and the RF equipment is deployed in a separate radio unit, referred to as a remote radio or active antenna unit in different cases, and then connected to a centralized processing unit, which is called the baseband unit (BBU), through fibers. Hence, the most urgent action, in our opinion, is to optimize the energy efficiency of the radio.

Furthermore, it is beneficial to analyze the distribution of energy consumption within a radio unit to achieve better energy efficiency optimization. However, this significantly depends on several implementation issues, such as the achievable power efficiency of the commercial PA and the resolution and sampling rate of the commercial analog-to-digital or digital-to-analog converters, and differs significantly for different production. However, the large power consumption of a 5G radio unit mainly comes from the massive antennas and associated TRX chains; hence, it is at least reasonable to decouple the energy consumption into two parts, such as the power consumption linear to the number of TRX chains and that irrelevant to the number of TRX chains [143]; in most cases, the former is dominant.

Another crucial observation is the energy consumption in the time domain. The traffic volume fluctuates significantly in different hours within a day; thus, the RU changes immensely. However, the variation of the average power consumption is extremely small. For example, the PRB usage is approximately 2% at 4 am, whereas the power consumption remains at 52% of the maximum value. Thus, the static power consumption, which is the power consumption of a transceiver for enabling fast activation, dominates the dynamic part that mainly comes from the PA and only exists when there is data transmission.

5.1.2 *Candidate energy-saving solutions for radio units*

There are several ways to optimize the energy efficiency of a radio unit, and these solutions can be mainly divided into two aspects: (1) optimizing the transceiver architecture and improving the energy efficiency of hardware components, e.g., improving the efficiency of PAs [144]. (2) improving the energy efficiency through suitable power control and signal processing for a given transceiver architecture. We will focus on the second aspect in this section. For the optimization of the network energy efficiency, the effort should be paid differently in different cases. According to the traffic demand, these three cases can be identified: the idle, light-load, and heavy-load cases.

(1) Optimization in the idle case. The idle case mainly represents the case in the evening when there is less traffic demand, and the main effort should be paid to reduce the static power consumption of the radio unit. The final goal is to turn off the transceiver and remove the static power totally when there is no traffic demand. However, this is unrealistic since the radio units must maintain some basic functions to enable a timely turning on to serve UEs with sudden access requests and traffic arrival. Specifically, there are some always-on common signals in 5G NR networks to guarantee quick activation and response in the idle case. For a typical 5G base station operating on a C-band, the total overhead for these always-on signals can increase to approximately 12%.

To this end, some researchers have proposed a heterogeneous network deployment, in which all the base stations except a hyper-base station can be turned off in the idle case, and only the hyper-base station can respond to the transmission of common signals/channels to enable UE access [145]. However, this design requires a thorough re-design and re-plan of the mobile network, which is challenging to achieve in practice. For the homogeneous network, the coverage of each base station is suitably optimized, and one base station is difficult to cover the UEs remote from it. It is difficult to turn off numerous base stations but preferably to turn off some carriers and retain one basic carrier with minimum bandwidth for the transmission of common signals/channels [146]. Another solution for the idle case is to re-design the cell access procedure. For example, a UE-triggered cell access can be used only when the UE has data for transmission; the UE sends a signal to wake up the base stations that would then transmit the common signals/channels to enable the UE to access the cell.

(2) Optimization in the light-load case. In the light-load case, there is a certain traffic demand, and the average PRB use is approximately 30%. In such cases, the most effective energy-saving method is to turn off some symbols and/or TRX chains suitably to reduce the static power consumption [146]. Such turning-off actions can be implemented either semi-statically, i.e., in large time-domain granularity, or dynamically, i.e., in small time-domain granularity. With large time-domain granularity, such as in the unit of seconds, the TRX chains are turned off, maintained, and muted for a long duration; hence, more circuit components can be muted to reduce the static power consumption to the lowest. However, this semi-static method has two non-negligible drawbacks; (1) the impact on the coverage of some common/control signals/channels; hence, some cell-edge UEs can receive these messages unsuccessfully. (2) This method cannot track the dynamic traffic arrival in time. The semi-static TRX muting adjustment scheme can lead to extra transmission delay on some slots with large traffic buffers or poor channel conditions.

In contrast, if a small time-domain granularity is used, some circuit components cannot be muted because of the long activation time, resulting in a relatively large static power consumption. However, the dynamic method is friendly to the network/user performance. For example, this dynamic turning off can be implemented in a per-slot approach to accommodate the channel/signal type, traffic buffer, and channel condition for data transmission; this can achieve the best power saving gains while ensuring that performance is not lost. However, if the pattern of the TRX chain changes dynamically on different slots, the link adaption becomes extremely difficult. Hence, it is desirable to design some enhanced CSI measure/feedback method to enable dynamic muting of TRX chains and accurate link adaption.

However, there is some data transmission in the light-load case; thus, the dynamic power consumption cannot be ignored. Another solution is to optimize the energy efficiency of data transmission. Based on the Shannon capacity formula, for a given amount of data, the total transmit power can be reduced by enlarging the resource amount, hence reducing the transmit power spectrum density (PSD). Thus, some researchers have proposed to spread the data transmission on as many resources as possible to reduce the total power consumption [147]. However, this method is unsuitable for cell-edge users because a reduced PSD would lead to a larger channel estimation error, which will significantly impact the achievable transmission rate. Another aspect that must be studied is the scaling laws between the resources and transmit power, specifically in a complicated case with unknown intra-cell and inter-cell interference.

(3) Optimization in the heavy-load case. In the heavy-load case, the traffic demand is extremely large, and the PRB use is higher than 50%. Furthermore, there is a small space for symbol/TRX chain muting, and the dynamic power consumption would be more dominant. Thus, it is more important to improve the energy efficiency of data transmission. Optimizing the PA efficiency is the most effective approach, and the most related methods are the low peak-to-average power ratio (PAPR) waveform [148] and PAPR-reduction methods [149].

5.1.3 *Other energy-saving solutions*

Currently, a BBU is linked to three to four radio units through the fronthaul links, and the power consumption of the fronthaul links is relatively small, which is approximately several tens of watt if the fronthaul links are based on fibers. However, if a more distributed deployment is adopted, where numerous radio units would be connected through additional fronthaul links, the energy consumption of fronthaul links becomes non-negligible. To this end, a lot of research has been conducted to optimize the energy efficiency of the distributed MIMO system based on the energy consumption in fronthaul links and the energy difference between the sleeping and active states of a radio unit [150]. Additionally, if the fronthaul links are based on radio waves, the energy consumption would increase significantly; thus, it is significant to optimize the E2E energy efficiency of data transmission by jointly considering the energy consumption in fronthaul links and the air interfaces.

Figure 31 shows that the energy consumption for digital signal processing is not dominant in the current base stations. However, as the transceiver antennas and transmission bandwidth increase, the signal processing at the baseband becomes significantly more complicated. Thus, it would be beneficial to investigate how to achieve a good balance between low-complexity/low-power signal processing and the achievable transmission rate to optimize the total energy efficiency in advance [151].

Finally, optimization reduces the carbon emissions and operating cost; thus, it is beneficial to introduce green energy for power supply in addition to the power grid. The green energy can be harvested locally in the base station; thus, the main problem is how to optimize the data transmission and energy buffer strategies in the time domain to minimize the extra power consumption from grids while guaranteeing the basic data requirements [152]. However, the green energy harvested at different places can be supplied back to the power grid, thus enabling a green energy flow in geometry. The main problem becomes more complicated as the energy and traffic must be matched in both time and spatial domains to achieve the best green energy consumption and data transmission [153].

5.2 Advanced wireless network architecture

The 5G-Advanced mobile network must meet diversified demands and require continuous innovation. The driving force behind the network architecture transformation includes the following factors [154]: complex networks integrating multiple services, standards, and site types; coordination of multi-connectivity technologies; on-demand deployment of service anchors; flexible orchestration of network functions; and shorter period of service deployment. To achieve this goal, 5G introduces new technologies and concepts such as virtualization technology, network slicing, service-based architecture (SBA) in the core network, and the wireless access network [155]. The 5G-Advanced wireless network architecture is elaborated as follows.

5.2.1 *Anchorless mobility*

An important KPI of a wireless network for a seamless mobility experience is 0 ms interruption. In a 5G wireless network, a dual-active protocol stack (DAPS) has been introduced to ensure 0 ms interruption using the dual user-plane anchor of both source and target cells. However, it is aimed only at URLLC traffic with single connectivity, i.e., without CA and multi-radio DC. Furthermore, adding more legs in the context of DAPS would significantly increase the complexity of the chipset, which increases the industrial bar. Thus, the 5G-Advanced wireless network should be designed for a seamless mobility experience in terms of guaranteed high data rate and low latency, and it should be friendly to implementation for commercialized terminals. Figure 32 shows that FSA mobility for dynamic serving cell change can be used to dynamically switch the serving cell via lower layer signaling without increasing the chipset complexity of the simultaneous transceiver, which creates the anchorless architecture within the mobility area. From stages 1 to 3, primary cell (PCell) and SCell can be changed to any candidate serving cells that have been pre-configured using L3 signaling. Lower-layer signaling is used to indicate the dynamic switch, thus eliminating the interruption time caused by L3 signaling, and the CA connectivity can be maintained during the mobility.

In addition to the air interface of 0 ms interruption, the E2E latency reduction is also considered as a requirement of the 5G-Advanced wireless network for URLLC and RTBC traffic. To avoid the latency over network interface caused by network path anchor relocation, the network paths can be pre-configured in the mobility area, and the UE can move without the network path anchor relocation, as shown in the Figure 33. Here, the data packet numbered 1, 2 and 3 can be transmitted over the appropriate candidate network path from the user plane function to the target distributed unit via the intermediate central unit so that 0 ms interruption over the network path can be achieved using the network path anchorless mobility.

With the increasing deployment of multipath quick user datagram protocol, Internet connection, and multipath transmission control protocol by the applications, the E2E path in the transportation layer can be flexibly established and discovered, and thus above network-centric mobility architecture can be further evolved to UE-centric mobility by moving the anchor out of the wireless network architectures with the benefit of a unified framework for any radio access technology (RAT) and scenarios.

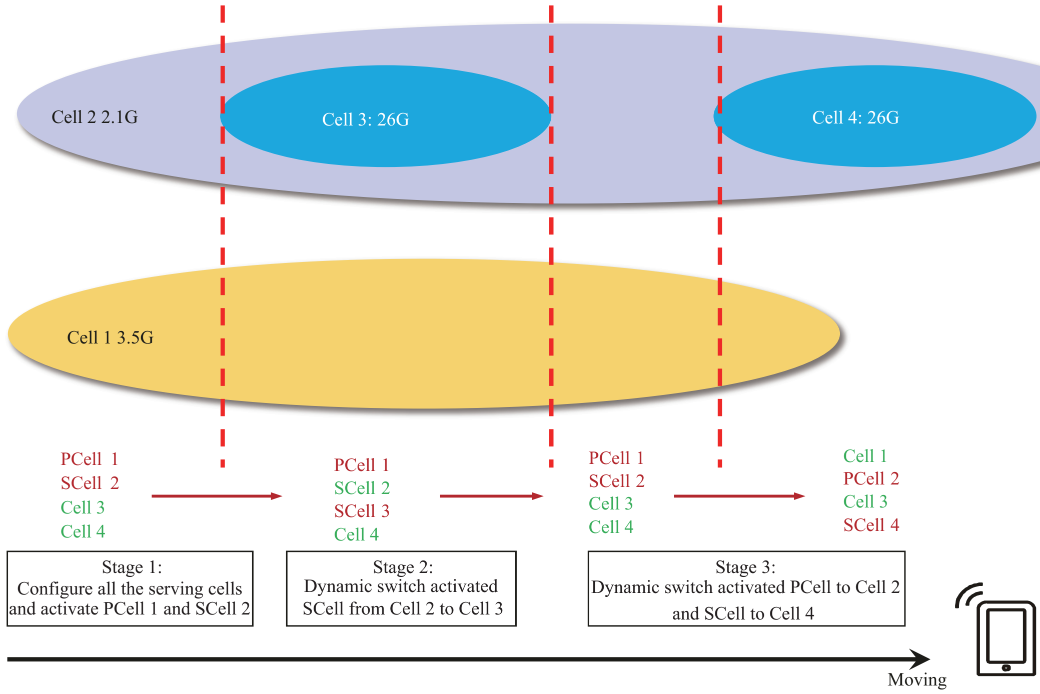

Figure 32 (Color online) Anchorless mobility with CA 0 ms interruption.

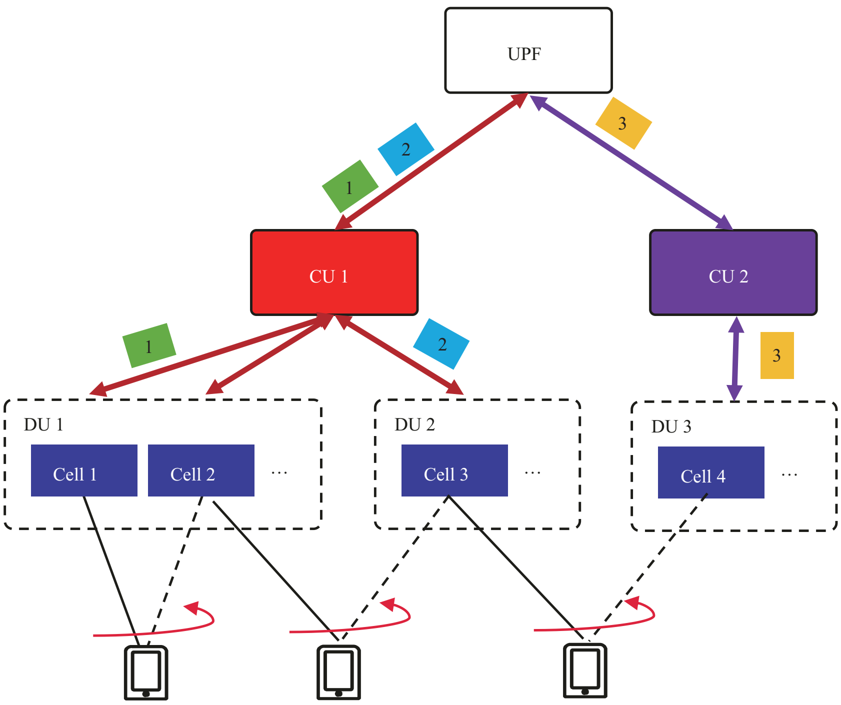

Figure 33 (Color online) Network path anchorless mobility across DU and CU.

5.2.2 *Service orchestration-oriented network*

To accommodate the diverse service requirement, the 5G-Advanced network can be orchestrated in the E2E service architecture to improve the 5G system performance for the targeted service. As shown in Figure 34, when the time-sensitive communication (TSC) traffic from multiple UEs arrives at RAN concurrently within some time slots, limited radio resources cannot accommodate these unexpected concurrent traffic arrivals. For TSC traffic, such scheduling delay is unacceptable, thus causing further resource inefficiency.

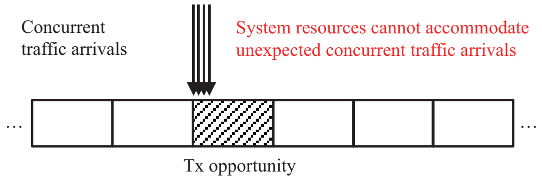

Figure 34 (Color online) An illustration of service concurrency at RAN side.

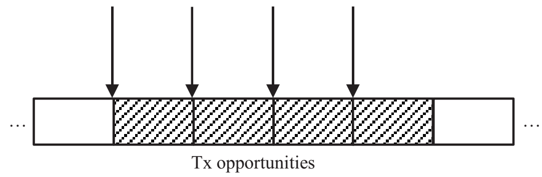

Figure 35 An illustration of traffic alignment at RAN side.

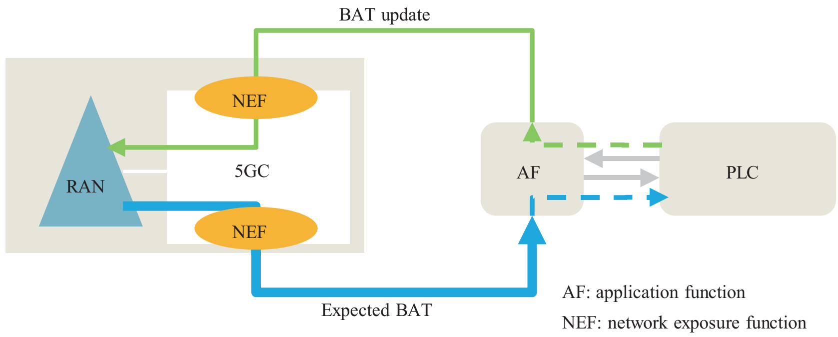

Figure 36 (Color online) Cross-layer service orchestration.

To resolve the above impacts, if the RAN reserved pattern can be coordinated with the TSC traffic pattern, it can be a promising solution to address this issue. However, if the burst arrival time (BAT) of different TSC UEs at the RAN side can be distributed more evenly, the RAN node will have sufficient capability and flexibility to schedule radio resources for traffic from different TSC UEs to guarantee their latency and reliability requirements. Figure 35 shows the desired effect of perfect alignment between TSC traffic arrivals and radio resources. Figure 36 illustrates the service orchestration-oriented network by coordinating with the service on the expected BAT from the RAN side to maximize the RAN resource usage.

Moreover, the wireless network architecture can be coordinated with the recent advances in the transportation layer techniques such as New IP by exploiting more possibilities of the cross-layer service orchestration to establish a tailed service-oriented wireless network architecture.

5.2.3 *Cooperative network*

Integrating the *cooperation* among the 5G nodes into the 5G network architecture would significantly improve the system scalability and maximize the performance in terms of reliability and throughput. UE backup and aggregation of multi-path can be considered for a cooperative network in the 5G-Advanced network architecture to fully explore the 5G resources to satisfy the stringent requirements of the RTBC and IIoT services.

(1) UE backup architecture. The E2E high reliability consists of not only the transmission link reliability but also the equipment reliability, which is equally essential in some industrial situations. For instance, a UE may break down after a long period of work because of software or hardware malfunction; when this occurs, the control system connected to the network via this UE will be unable to maintain stability, thus triggering preventive measures such as production line breakdown that will cause significant danger to E2E traffic reliability. Figure 37 illustrates a potential high-efficient UE backup network architecture, where UE 2 can quickly and efficiently take over UE 1's transmission when it breaks down. In normal time, the traffic data is duplicated to both UEs while only UE 1 performs the transmission. Thus, only one set of UE power resources, access-link resource, transport resource, and processing resource is required. As URLLC has strict latency and reliability requirements for transmission, the taking-over procedure should be completed within an extremely short time (below survival time level) and without data loss.

(2) UE aggregation architecture. Figure 38 shows that the existing SL-based relay supports one path for data transmission at any time. In some scenarios, the remote UE is in coverage, or it can see two or more relay UE(s) with good link quality when it is out of coverage. In these cases, the multi-path

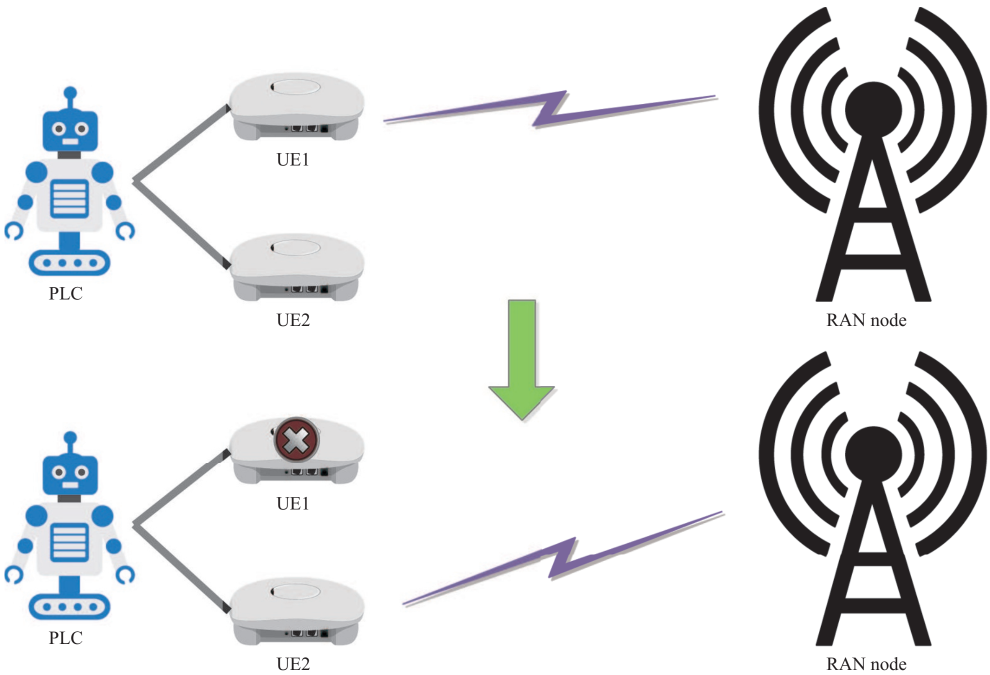

Figure 37 (Color online) High-efficient UE backup architecture.

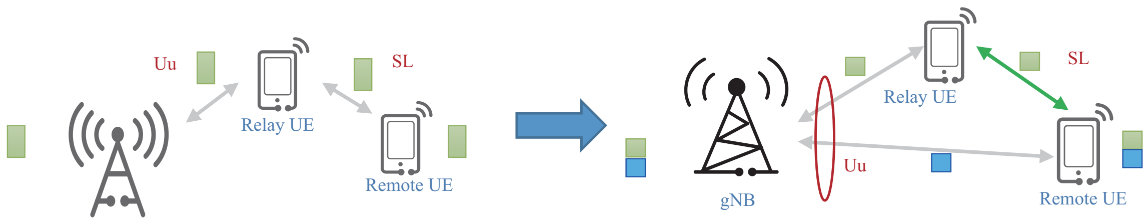

Figure 38 (Color online) UE aggregation of multi-path cooperation.

connection could be established between the remote UE and the gNB. These paths include both direct and indirect paths. The data of remote UE could be split/duplicated at the PDCP layer, for example, and distributed/transmitted over the multi-path. At the gNB side, the data received from the multi-path could be aggregated at the PDCP layer. The protocol stack used is similar to that of the DC. With multi-path connection and transmission of remote UE data by both relay and remote UEs, it is equivalent to boosting the transmit power of the remote UE, thus overcoming the bottleneck of uplink transmitted power limitation such as increasing the transmitted power from 23 to 26 dBm. Additionally, other benefits such as increased diversity/multiplexing could be exploited, as mentioned in Subsection 3.4.1.

5.2.4 *Intelligent and secure network*

5G-Advanced wireless networks should have a scalable network architecture that uses fog (or edge) computing technologies to distribute computing, storage, communication, and network functions closer to end-users; it should intelligently provide automated network management, flexible, and efficient resource allocation. For example, to better support real-time broadband interactive services such as XR and Tactile, applications or rendering tasks with a large amount of computing and storage must be placed as close to end-users as possible. Thus, wireless caching, fog computing, and mobile edge computing (MEC) play an important role in the 5G-Advanced wireless network architecture; the traditional on-demand based content delivery network should be developed under the control of the operator to further improve the efficiency of content distribution and resolve the congestion at the backhaul link. The 5G-Advanced wireless network architecture can be further improved using the integration of smart caching/MEC tools, AI, and block chain, which can be summarized in Table 3.

Table 3 Intelligent and secure network architecture

Technology directions	Corresponding scenarios	Description
Smart caching-integrated network	eMBB RTBC	5G-Advanced network architecture can be deployed with smart caching close to the RAN side and/or in MEC server for reducing the long round-trip of on-demand content delivery.
AI-integrated network	eMBB RTBC MTC	5G-Advanced network architecture can offer the AI computing through AI training, data collection and statistics through one or more gNBs for network planning optimization.
Block chain-integrated network	IIOT V2X eMBB	5G-Advanced network architecture can establish the secure communication environment where each 5G node can be participated in the blockchain establishment and secure information delivery.

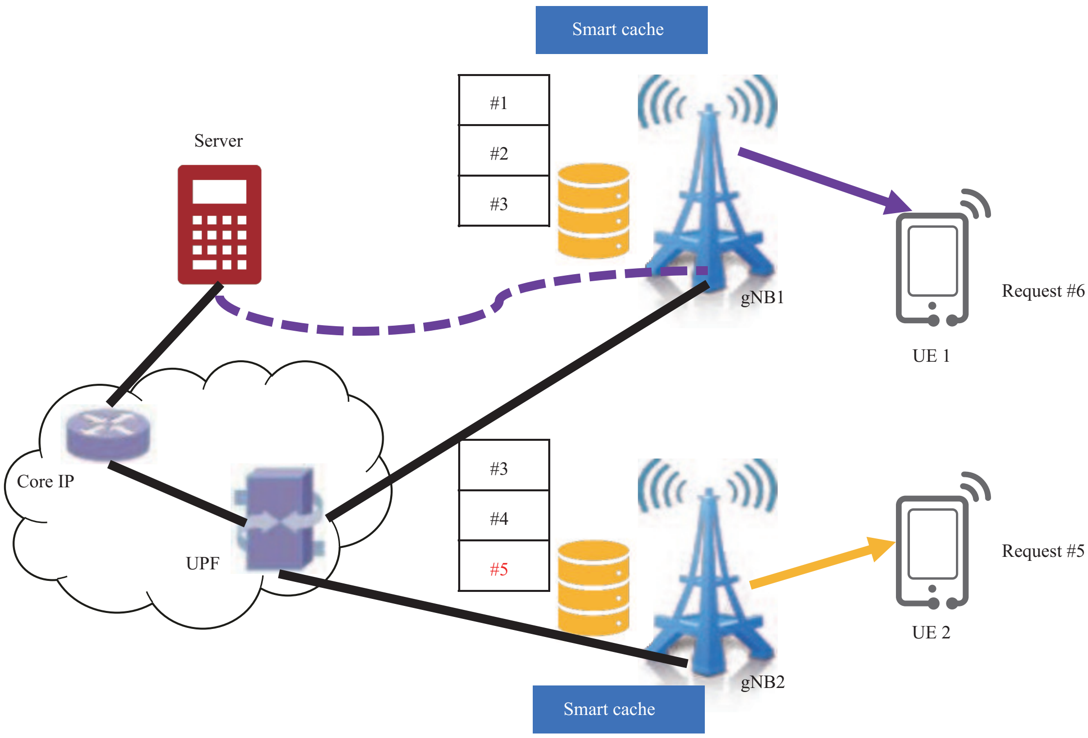

Figure 39 (Color online) Smart caching-integrated network architecture.

(1) Smart caching-integrated network. Figure 39 shows that caching can be implemented close to the RAN side, and the gNB can decide to distribute the content based on the caching policy and the UE's request. For instance, UE 1 requests file #6, but the serving gNB does not store the corresponding file in the caching; thus, gNB 1 will further initiate the content delivery from the server and then delivers to the UE 1. As file #5 is located in the cache associated with gNB 2, gNB 2 can directly deliver the requested file #5 to the UE 2. Note that the smart caching can also be deployed in the MEC server, and the MEC server can be responsible for caching update and policy distribution among multiple gNBs; thus, the MEC with smart caching can be further investigated in the 5G-Advanced network architecture [156].

(2) AI-integrated network. The increase of spectrum efficiency due to the advanced radio technology poses a huge challenge to wireless network planning, management, and operation. Traditional approaches of human experience and simple algorithms have drawbacks such as time-consuming, high cost, and poor adaptation to the diverse RRM policies. The use of AI in wireless networks has huge potential for the efficient operation of future wireless networks. The AI-aided wireless network can be used to cope with the complex and dynamic wireless environment; it can be used to overcome the drawbacks of traditional communication algorithms with the aid of AI algorithm running over the 5G system by collecting and analyzing data through either one or multiple gNBs [157].

(3) Block chain-integrated network. As 5G service and network architecture are undergoing significant reshaping, more diversified service scenarios and new network architectures will drive the transformation of 5G security architecture. This will promote industry partners to build a new security ecosystem to meet the digital transformation requirements of various 5G industries. In the future, if security protection and privacy are considered, the features and strength of the 5G network can be improved. Further, 5G-Advanced challenges can be handled easily with the capability of handling an increasing number of heterogeneous sources and devices; it can also be used to protect users' data and organizations' software, detection, and imaging tools [158]. To support diversified applications using various QoS requirements in various network deployment scenarios, blockchain technology is used to securely and efficiently share computing, storage, and communication resources among different communication standards and network nodes, and it can be applied to V2X, IIoT, and dynamic spectrum sharing scenarios. Consider the V2X scenario, where each vehicle participates in the blockchain-based secure authentication system by exchanging beacons, and the trustful V2X can be used for networking. In the secure communication environment, a security-sensitive message, such as sensing and security announcement, can be authenticated and integrity protected by each participating vehicle.

In summary, 5G-Advanced wireless networks should provide ubiquitous broadband connectivity, flexible, real-time, and distributed data processing capabilities. Deploying an AI-integrated wireless architecture based on fog (or edge) computing and the blockchain technology that allows multiple distributed nodes to process data on the network will significantly enable UCBC, RTBC, HCS, and automated network management.

6 Conclusion

The development history of 2G, 3G, and 4G over the past 30 years shows that, even in each generation, continuous evolution is crucial for unlocking the full potential of the technology and advancing the industry. This also applies to 5G, which is now accelerating its evolution to 5G-Advanced as we continue its commercial rollout and become more connected. Furthermore, 5G-Advanced is an evolution of 5G with more application expansion. Additionally, to continuously improve the three standard scenarios of eMBB, mMTC, and URLLC defined by the ITU for 5G, 5G-Advanced will introduce three new scenarios such as UCBC, RTBC, and HCS to support everything that is connected and intelligent.

In this study, the key driving forces, requirements, usage scenarios, and capabilities of 5G-Advanced have been proposed. Furthermore, promising technological directions have been provided, and the enabling technologies have been investigated. Because of space constraints, this study does not cover other promising application scenarios and related technological directions such as mobile IAB (aka. vehicle-mounted relay), NR multicast broadcast services, and non-terrestrial networks, which include satellite communication networks, high altitude platform systems, air-to-ground networks, and unmanned aerial vehicles. We hope that this study will provide academic researchers and industry experts with some insightful ideas on the evolution of 5G toward 5G-Advanced. Furthermore, we also look forward to working with all partners to successfully achieve 5G-Advanced to build a better and more intelligent world together.

References

1 Chen Y, Zhu P Y, He G N, et al. From connected people, connected things, to connected intelligence. In: Proceedings of the 2nd 6G Wireless Summit (6G SUMMIT), 2020. 1–7
2 Ghosh A, Maeder A, Baker M, et al. 5G evolution: a view on 5G cellular technology beyond 3GPP release 15. IEEE Access, 2019, 7: 127639
3 Bertenyi B. 5G evolution: what's next? IEEE Wirel Commun, 2021, 28: 4–8
4 Release 18 comes into view. https://www.3gpp.org/release18
5 Levy K. Huawei's perspective on IoT & Industry 4.0 in the 5G era. 2021. https://www.gsma.com/iot/wp-content/uploads/2021/07/Mobile-IoT-Summit-2-Huawei_IoT-and-Industry-4.0-in-the-5G-Era.pdf
6 Zhu P Y. 5G to 5.5G: wireless innovation is an endless frontier. 2021. https://www.huawei.com/en/news/2021/3/zhupeiying-ieee-wcnc2021
7 Capacity Media. 5G is so last year: now we're moving to 5G-Advanced. 2021. https://www.capacitymedia.com/articles/3828674/5g-is-so-last-year-now-were-moving-to-5g-advanced
8 Huawei's Global Industry Vision Report. 10 trends for 2025: touching the intelligent world. 2019. https://www.huawei.com/minisite/giv/Files/whitepaper_en_2019.pdf

9 GTI. 5G wireless evolution white paper: towards a sustainable 5G. 2021. https://www-file.huawei.com/-/media/CORP2020/pdf/event/1/5G_Advanced_Technology_Evolution_from_a_Network_Perspective_2021_en.pdf

10 Wang T. Defining 5.5G for a better, intelligent world. 2020. https://www.huawei.com/minisite/mbbf2020/en/

11 3GPP. 3GPP/PCG#46-e draft report V2.0. 2021. https://www.3gpp.org/ftp/PCG/PCG_46/Report/PCG47_02.zip

12 Strategy Advisory Board. 6G wireless: a new strategic vision. 2020. https://www.surrey.ac.uk/sites/default/files/2020-11/6g-wireless-a-new-strategic-vision-paper.pdf

13 Zhang P, Niu K, Tian H, et al. Technology prospect of 6G mobile communications. J Commun, 2019, 40: 141–148

14 You X H, Wang C X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301

15 IMT-2030 (6G) Promotion Group. 6G vision and candidate technologies. 2021. http://www.caict.ac.cn/english/news/202106/P020210608349616163475.pdf

16 Tong W, Zhu P Y. 6G: the Next Horizon: from Connected People and Things to Connected Intelligence. Cambridge: Cambridge University Press, 2021

17 Yuan Y F, Zhao Y J, Zong B Q, et al. Potential key technologies for 6G mobile communications. Sci China Inf Sci, 2020, 63: 183301

18 3GPP. RWS-210435: Rel-18 overview. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_AHs/2021_06_RAN_Rel18_WS/Docs/RWS-210435.zip

19 Feng D Q, Lai L F, Luo J J, et al. Ultra-reliable and low-latency communications: applications, opportunities and challenges. Sci China Inf Sci, 2021, 64: 120301

20 Tong W. 5G continuous evolution for building engine of all industry digitalization. 2021. https://www.huawei.com/en/news/2021/2/5g-evolution-engine-digitalization

21 Wan L, Anthony S, Liu J H, et al. 5G System Design: an End to End Perspective. Berlin: Springer, 2020

22 Navarro-Ortiz J, Romero-Diaz P, Sendra S, et al. A survey on 5G usage scenarios and traffic models. IEEE Commun Surv Tut, 2020, 22: 905–929

23 Onggosanusi E, Rahman M S, Guo L, et al. Modular and high-resolution channel state information and beam management for 5G new radio. IEEE Commun Mag, 2018, 56: 48–55

24 3GPP. R1-1812242: discussion on CSI enhancement. 2018. https://www.3gpp.org/ftp/tsg_ran/wg1_rl1/TSGR1_95/Docs/R1-1812242.zip

25 3GPP. R1-2006414: discussion on field measurement and evaluation assumptions for FDD CSI enhancements in Rel-17. 2020. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_102-e/Docs/R1-2006414.zip

26 3GPP. R1-2007592: discussion on CSI enhancements for Rel-17. 2020. https://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_103-e/Docs/R1-2007592.zip

27 Yin H F, Wang H Q, Liu Y Z, et al. Addressing the curse of mobility in massive MIMO with prony-based angular-delay domain channel predictions. IEEE J Sel Areas Commun, 2020, 38: 2903–2917

28 3GPP. R1-2007591: discussion on SRS enhancements for Rel-17. 2020. https://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_103-e/Docs/R1-2007591.zip

29 Björnson E, Sanguinetti L, Wymeersch H, et al. Massive MIMO is a reality-what is next?: Five promising research directions for antenna arrays. Digit Signal Process, 2019, 94: 3–20

30 Martínez À O, Nielsen J O, de Carvalho E, et al. An experimental study of massive MIMO properties in 5G scenarios. IEEE Trans Antennas Propagat, 2018, 66: 7206–7215

31 Martínez À O, de Carvalho E, Nielsen J Ø. Towards very large aperture massive MIMO: a measurement based study. In: Proceedings of EEE Globecom Workshops, 2014, 281–286

32 Payami S, Tufvesson F. Channel measurements and analysis for very large array systems at 2.6 GHz. In: Proceedings of the 6th European Conference on Antennas and Propagation (EUCAP), 2012, Prague. 433–437

33 Gao X, Edfors O, Rusek F, et al. Massive MIMO performance evaluation based on measured propagation data. IEEE Trans Wirel Commun, 2015, 14: 3899–3911

34 Han Y, Jin S, Wen C K, et al. Channel estimation for extremely large-scale massive MIMO systems. IEEE Wirel Commun Lett, 2020, 9: 633–637

35 de Carvalho E, Ali A, Amiri A, et al. Non-stationarities in extra-large-scale massive MIMO. IEEE Wirel Commun, 2020, 27: 74–80

36 Li K, Sharan R R, Chen Y, et al. Decentralized baseband processing for massive MU-MIMO systems. IEEE J Emerg Sel Top Circ Syst, 2017, 7: 491–507

37 Jeon C, Li K, Cavallaro J R, et al. Decentralized equalization with feedforward architectures for massive MU-MIMO. IEEE Trans Signal Process, 2019, 67: 4418–4432

38 Li K, Jeon C, Cavallaro J R, et al. Feedforward architectures for decentralized precoding in massive MU-MIMO systems. In: Proceedings of the 52nd Asilomar Conference on Signals, Systems, and Computers, 2018. 1659–1665

39 Sanchez J R, Rusek F, Edfors O, et al. Decentralized massive MIMO processing exploring daisy-chain architecture and recursive algorithms. IEEE Trans Signal Process, 2020, 68: 687–700

40 Shaik Z H, Björnson E, Larsson E G. MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes. 2020. arXiv:2012.13928

41 Amiri A, Angjelichinoski M, de Carvalho E, et al. Extremely large aperture massive MIMO: low complexity receiver architectures. In: Proceedings of IEEE Globecom Workshops, 2018, Abu Dhabi. 1–6

42 3GPP. Study on channel model for frequencies from 0.5 to 100 GHz. TR 38.901. https://www.3gpp.org/ftp//Specs/archive/38_series/38.901/38901-g10.zip

43 Flordelis J, Li X, Edfors O, et al. Massive MIMO extensions to the COST 2100 channel model: modeling and validation. IEEE Trans Wirel Commun, 2020, 19: 380–394

44 You X H, Wang D M, Wang J Z. Distributed MIMO and Cell-Free Mobile Communication. Beijing: Science Press, 2019

45 Yoo I, Imani M F, Sleasman T, et al. Enhancing capacity of spatial multiplexing systems using reconfigurable cavity-backed metasurface antennas in clustered MIMO channels. IEEE Trans Commun, 2019, 67: 1070–1084

46 Bahceci I, Hasan M, Duman T M, et al. Efficient channel estimation for reconfigurable MIMO antennas: training techniques and performance analysis. IEEE Trans Wirel Commun, 2017, 16: 565–580

47 Hasan M, Bahceci I, Cetiner B A. Downlink multi-user MIMO transmission for radiation pattern reconfigurable antenna systems. IEEE Trans Wirel Commun, 2018, 17: 6448–6463

48 Zhao T C, Li M, Ditzler G. Online reconfigurable antenna state selection based on Thompson sampling. In: Proceedings of

International Conference on Computing, Networking and Communications, 2019, Honolulu. 888–893
49 Chen X H, Liu A, Cai Y L, et al. Randomized two-timescale hybrid precoding for downlink multicell massive MIMO systems. IEEE Trans Signal Process, 2019, 67: 4152–4167
50 Shyianov V, Akrout M, Bellili F, et al. Achievable rate with antenna size constraint: Shannon meets Chu and Bode. 2020. arXiv:2011.05529
51 Pizzo A, Marzetta T, Sanguinetti L. Spatially-stationary model for holographic MIMO small-scale fading. 2019. arXiv:1911.04853
52 Haneda K, Gustafson C, Wyne S. 60 GHz spatial radio transmission: multiplexing or beamforming? IEEE Trans Antennas Propagat, 2013, 61: 5735–5743
53 Shalaginov M Y, An S, Zhang Y F, et al. Reconfigurable all-dielectric metalens with diffraction-limited performance. Nature Commun, 2021, 12: 1–8
54 Hemadeh I A, Xiao P, Kabiri Y, et al. Polarization modulation design for reduced RF chain wireless. IEEE Trans Commun, 2020, 68: 3890–3907
55 Williams R J, Carvalho E D, Marzetta T L. A communication model for large intelligent surfaces. 2019. arXiv:1912.06644
56 Sarkar D, Mikki S, Antar Y M M. Engineering the eigenspace structure of massive MIMO links through frequency-selective surfaces. Antennas Wirel Propag Lett, 2019, 18: 2701–2705
57 Liu R, Wu Q, Renzo M D, et al. A path to smart radio environments: an industrial viewpoint on reconfigurable intelligent surfaces. 2021. arXiv:2104.14985
58 di Renzo M, Zappone A, Debbah M, et al. Smart radio environments empowered by reconfigurable intelligent surfaces: how it works, state of research, and the road ahead. IEEE J Sel Areas Commun, 2020, 38: 2450–2525
59 Chen W C, Wen C K, Li X, et al. Adaptive bit partitioning for reconfigurable intelligent surface assisted FDD systems with limited feedback. 2020. arXiv:2011.14748
60 Chen J, Liang Y C, Cheng H, et al. Channel estimation for reconfigurable intelligent surface aided multi-user MIMO systems. 2019. arXiv:1912.03619
61 Alwazani H, Nadeem Q-U-A, Chaaban A. Channel estimation for distributed intelligent reflecting surfaces assisted multi-user MISO systems. In: Proceedings of IEEE Globecom Workshops, 2020, Taipei. 1–6
62 Cai W H, Liu R, Liu Y, et al. Intelligent reflecting surface assisted multi-cell multi-band wireless networks. In: Proceedings of IEEE Wireless Communications and Networking Conference, 2021, Nanjing. 1–6
63 Mhanna E, Mohamad A, Debbah M, et al. Distributed stochastic phase-shift optimization in a RIS-assisted cellular network. In: Proceedings of IEEE Wireless Communications and Networking Conference, 2021, Nanjing. 1–6
64 Liu R, Li M, Liu Q, et al. Joint symbol-level precoding and reflecting designs for IRS-enhanced MU-MISO systems. IEEE Trans Wirel Commun, 2021, 20: 798–811
65 Zhang S H, Zhang H L, Di B Y, et al. Intelligent omni-surface: ubiquitous wireless transmission by reflective-transmissive metasurface. 2020. arXiv:2011.00765
66 3GPP. QoE parameters and metrics relevant to the virtual reality (VR) user experience. TR 26.929. https://www.3gpp.org/ftp//Specs/archive/26_series/26.929/26929-g10.zip
67 Dou S Y, Liao S R, Wu J, et al. XR quality index: evaluating RAN transmission quality for XR services over 5G and beyond. In: Proceedings of the 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications, 2021. 1–6
68 3GPP. Extended reality (XR) in 5G. TR 26.928. https://www.3gpp.org/ftp//Specs/archive/26_series/26.928/26928-g10.zip
69 3GPP. R1-2102308: reply LS on new standardized 5QIs for 5G-AIS (Advanced Interactive Services). 2021. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_104b-e/LS/Incoming/R1-2102308.zip
70 3GPP. Chairman notes, RAN1#104b-e, April 12–April 20, 2021. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_104b-e/Inbox/Chair_notes/Chair's%20Notes%20RAN1%23104b-e%20final.zip
71 3GPP. R1-2105521: initial evaluation results for XR and Cloud Gaming. 2021. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_105-e/Docs/R1-2105521.zip
72 Chen E K, Dou S Y, Wang S, et al. Frame-level integrated transmission for extended reality over 5G and beyond. In: Proceedings of IEEE Global Communications Conference, 2021, Madrid. 1–6
73 Reed I S, Solomon G. Polynomial codes over certain finite fields. J Soc Indust Appl Math, 1960, 8: 300–304
74 Ho T, Medard M, Koetter R, et al. A random linear network coding approach to multicast. IEEE Trans Inform Theor, 2006, 52: 4413–4430
75 Internet Engineering Task Force (IETF). RFC5053: raptor forward error correction scheme for object delivery. 2007. https://www.rfc-editor.org/rfc/pdfrfc/rfc5053.txt.pdf
76 Gluesing-Luerssen H, Rosenthal J, Smarandache R. Strongly-MDS convolutional codes. IEEE Trans Inform Theor, 2006, 52: 584–598
77 Tomas V, Rosenthal J, Smarandache R. Decoding of convolutional codes over the erasure channel. IEEE Trans Inform Theor, 2012, 58: 90–108
78 Badr A, Lui D, Khisti A. Streaming codes for multicast over burst erasure channels. IEEE Trans Inform Theor, 2015, 61: 4181–4208
79 Kong D J, Xia X G, Liu P, et al. MMSE channel estimation for two-port demodulation reference signals in new radio. Sci China Inf Sci, 2021, 64: 169303
80 Ma C X, Liu R K, Liao S, et al. User cooperation scheduling in cellular systems. In: Proceedings of IEEE Globecom Workshops, 2020, Taipei. 1–6
81 Singh B, Tirkkonen O, Li Z, et al. Contention-based access for ultra-reliable low latency uplink transmissions. IEEE Wirel Commun Lett, 2018, 7: 182–185
82 Kotaba R, Manchon C N, Balercia T, et al. Uplink transmissions in URLLC systems with shared diversity resources. IEEE Wirel Commun Lett, 2018, 7: 590–593
83 Elayoubi S E, Brown P, Deghel M, et al. Radio resource allocation and retransmission schemes for URLLC over 5G networks. IEEE J Sel Areas Commun, 2019, 37: 896–904
84 Combes R, Elayoubi S E, Varela T, et al. Optimal retransmission policies for ultra-reliable low latency communications with delayed feedback. In: Proceedings of IEEE Global Communications Conference, 2019, Waikoloa. 1–6
85 Zhao S, Wang Y, Xie Y, et al. Joint time-frequency diversity based uplink grant-free transmission scheme for URLLC. In: Proceedings of International Conference on Wireless Communications and Signal Processing, 2019, Xi'an. 1–6
86 Esswie A A, Pedersen K I, Mogensen P E. Preemption-aware rank offloading scheduling for latency critical communications

in 5G networks. In: Proceedings of the 89th Vehicular Technology Conference, 2019, Kuala Lumpur. 1–6
87 Esswie A A, Pedersen K I. Capacity optimization of spatial preemptive scheduling for joint URLLC-eMBB traffic in 5G new radio. In: Proceedings of IEEE Globecom Workshops, 2018, Abu Dhabi. 1–6
88 Karimi A, Pedersen K I, Mahmood N H. Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G. In: Proceedings of the 89th Vehicular Technology Conference, 2019, Kuala Lumpur. 1–6
89 3GPP. RWS-210442: complementary TDD and URLLC enhancements for NR. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_AHs/2021_06_RAN_Rel18_WS/Docs/RWS-210442.zip
90 3GPP. RP-210918: WID on support of reduced capability NR devices. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_91e/Docs/RP-210918.zip
91 3GPP. Study on support of reduced capability NR devices. TR 38.875. https://www.3gpp.org/ftp/Specs/archive/38_series/38.875/38875-h00.zip
92 Liu Q, Sun S L, Yuan X G, et al. Ambient backscatter communication-based smart 5G IoT network. J Wirel Com Network, 2021, 2021: 34
93 3GPP. RWS-210453: passive IoT for 5G advanced. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_AHs/2021_06_RAN_Rel18_WS/Docs/RWS-210453.zip
94 Mao Q, Hu F, Hao Q. Deep learning for intelligent wireless networks: a comprehensive survey. IEEE Commun Surv Tut, 2018, 20: 2595–2621
95 3GPP. RP-201620: study on enhancement for data collection for NR and ENDC. 2020. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_89e/Docs/RP-201620.zip
96 3GPP. RP-213599: study on artificial intelligence (AI)/machine learning (ML) for NR air interface. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_94e/Docs/RP-213599.zip
97 Sohrabi F, Chen Z, Yu W. Deep active learning approach to adaptive beamforming for mmWave initial alignment. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, Toronto. 4940–4944
98 Wen C K, Shih W T, Jin S. Deep learning for massive MIMO CSI feedback. IEEE Wirel Commun Lett, 2018, 7: 748–751
99 Zhu J K, Chai M Y, Zhou W Y. Three-three-three network architecture and learning optimization mechanism for B5G/6G. J Commun, 2021, 42: 62–75
100 Zhu J K, Zhao M, Zhang S H, et al. Exploring the road to 6G: ABC-foundation for intelligent mobile networks. China Commun, 2020, 17: 51–67
101 Soltani M, Pourahmadi V, Mirzaei A, et al. Deep learning-based channel estimation. IEEE Commun Lett, 2019, 23: 652–655
102 Borgerding M, Schniter P, Rangan S. AMP-inspired deep networks for sparse linear inverse problems. IEEE Trans Signal Process, 2017, 65: 4293–4308
103 Cui Y, Li S, Zhang W. Jointly sparse signal recovery and support recovery via deep learning with applications in MIMO-based grant-free random access. IEEE J Sel Areas Commun, 2021, 39: 788–803
104 del Peral-Rosado J A, Raulefs R, Lopez-Salcedo J A, et al. Survey of cellular mobile radio localization methods: from 1G to 5G. IEEE Commun Surv Tut, 2018, 20: 1124–1148
105 Keating R, Säily M, Hulkkonen J, et al. Overview of positioning in 5G new radio. In: Proceedings of the 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, 2019. 320–324
106 Zhu P Y. Integrated sensing and communication for 6G opportunities and challenges. IEEE ComSoc ISAC-ETI Webinar Series, 2020. https://www.youtube.com/watch?v=V50CGOEVdEo
107 Cui Y H, Liu F, Jin X J, et al. Integrating sensing and communications for ubiquitous IoT: applications, trends and challenges. 2021. arXiv:2104.11457
108 IEEE Communications Society. Integrated sensing and communication emerging technology initiative. https://isac.committees.comsoc.org
109 3GPP. RWS-210438: NR FR2 enhancements. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_AHs/2021_06_RAN_Rel18_WS/Docs/RWS-210438.zip
110 Li X, Ma N, Tang Q, et al. Buffered DL/UL traffic ratio sensing cell clustering for interference mitigation in LTE TDD system. In: Proceedings of IEEE Wireless Communications and Networking Conference, 2018, Barcelona. 1–6
111 Nasreddine J, Hassan S E H. Interference mitigation and traffic adaptation using cell clustering for LTE-TDD systems. In: Proceedings of IEEE International Multidisciplinary Conference on Engineering Technology, 2016, Beirut. 155–159
112 Lukowa A, Venkatasubramanian V. Centralized UL/DL resource allocation for flexible TDD systems with interference cancellation. IEEE Trans Veh Technol, 2019, 68: 2443–2458
113 Kim H, Lee K, Wang H, et al. Cross link interference mitigation schemes in dynamic TDD systems. In: Proceedings of the 90th Vehicular Technology Conference, 2019, Honolulu. 1–5
114 Hiltunen K, Matinmikko-Blue M. Interference control mechanism for 5G indoor micro operators utilizing dynamic TDD. In: Proceedings of the 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, 2018, Bologna. 1–7
115 Ding M, Perez D L, Vasilakos A V, et al. Dynamic TDD transmissions in homogeneous small cell networks. In: Proceedings of IEEE International Conference on Communications Workshops, 2014, Sydney. 616–621
116 de Olivindo Cavalcante E, Fodor G, Silva Y C B, et al. Distributed beamforming in dynamic TDD MIMO networks with BS to BS interference constraints. IEEE Wirel Commun Lett, 2018, 7: 788–791
117 Huang Y, Jalaian B, Russell S, et al. Reaping the benefits of dynamic TDD in massive MIMO. IEEE Syst J, 2019, 13: 117–124
118 Lee K, Park Y, Na M, et al. Aligned reverse frame structure for interference mitigation in dynamic TDD systems. IEEE Trans Wirel Commun, 2017, 16: 6967–6978
119 Ardah K, Fodor G, Silva Y C B, et al. A novel cell reconfiguration technique for dynamic TDD wireless networks. IEEE Wirel Commun Lett, 2018, 7: 320–323
120 Guo S Z, Hou X L, Wang H N. Dynamic TDD and interference management towards 5G. In: Proceedings of IEEE Wireless Communications and Networking Conference, 2018, Barcelona. 1–6
121 Majeed E, Iwelski S, Bai Z J, et al. Advanced receiver design for interfering small cell deployments in LTE networks. In: Proceedings of IEEE Conference on Standards for Communications and Networking, 2015, Tokyo. 294–299
122 Takeda K, Harada H, Sano Y, et al. Higher order modulation, small cell discovery and interference cancellation technologies in LTE-Advanced release 12. NTT DOCOMO Tech J, 2015, 17: 47–55
123 Ding M, Lopez-Perez D, Xue R, et al. On dynamic time-division-duplex transmissions for small-cell networks. IEEE Trans Veh Technol, 2016, 65: 8933–8951

124 Yang J L, Zhang Y, Zhang D S, et al. Highly selective filter for suppressing interference of 5G signals to C-band satellite receiver. In: Proceedings of International Wireless Communications and Mobile Computing, 2020, Limassol. 24–27

125 Peccarelli N, Irazoqui R, Fulton C. Mitigation of interferers and nonlinear spurious products for digital array and MIMO systems. In: Proceedings of IEEE MTT-S International Microwave Symposium (IMS), 2019, Boston. 1233–1236

126 Harvanek M, Marsalek R, Kral J, et al. Adjacent channel interference cancellation in FDM transmissions. IEEE Trans Circ Syst I, 2020, 67: 5417–5428

127 Ali Z, Duel-Hallen A, Hallen H. Early warning of mmWave signal blockage and AoA transition using sub-6 GHz observations. IEEE Commun Lett, 2020, 24: 207–211

128 Dupleich D, Mueller R, Landmann M, et al. Multi-band propagation and radio channel characterization in street canyon scenarios for 5G and beyond. IEEE Access, 2019, 7: 160385

129 Mateo P J, Pizarro A B, Ludant N, et al. A comprehensive study of low frequency and high frequency channel correlation. In: Proceedings of International Conference on Computing, Networking and Communications, 2019, Honolulu. 876–882

130 Ali A, Gonzalez-Prelcic N, Heath R W. Spatial covariance estimation for millimeter wave hybrid systems using out-of-band information. IEEE Trans Wirel Commun, 2019, 18: 5471–5485

131 Han Y, Hsu T H, Wen C K, et al. Efficient downlink channel reconstruction for FDD multi-antenna systems. IEEE Trans Wirel Commun, 2019, 18: 3161–3176

132 Xu M, Zhang S, Zhong C J, et al. Ordinary differential equation-based CNN for channel extrapolation over RIS-assisted communication. IEEE Commun Lett, 2021, 25: 1921–1925

133 Arnold M, Dörner S, Cammerer S, et al. Towards practical FDD massive MIMO: CSI extrapolation driven by deep learning and actual channel measurements. In: Proceedings of the 53rd Asilomar Conference on Signals, Systems, and Computers, 2019, Pacific Grove. 1972–1976

134 Shen Y F, Shi Y M, Zhang J, et al. A graph neural network approach for scalable wireless power control. In: Proceedings of IEEE Globecom Workshops, 2019, Waikoloa. 1–6

135 Eisen M, Ribeiro A. Optimal wireless resource allocation with random edge graph neural networks. IEEE Trans Signal Process, 2020, 68: 2977–2991

136 Haj-Ali A, Ahmed N K, Willke T, et al. A view on deep reinforcement learning in system optimization. 2019. arXiv:1908.01275

137 GSMA. 5G energy efficiencies: green is the new black. 2020. https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=54165956&file=241120-5G-energy.pdf

138 ITU-T. ICT industry to reduce greenhouse gas emissions by 45 per cent by 2030. 2020. https://www.itu.int/en/mediacentre/Pages/PR04-2020-ICT-industry-to-reduce-greenhouse-gas-emissions-by-45-percent-by-2030.aspx

139 CMCC. The 2020 annual report. https://www.chinamobileltd.com/en/ir/reports/ar2020/2020_20f.pdf

140 ITU-T. Technical report on smart energy saving of 5G base station: based on AI and other emerging technologies to forecast and optimize the management of 5G wireless network energy consumption. 2021. https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Documents/TR-D.WG3_02-Smart%20Energy%20Saving%20of%205G%20Base%20Station%20Based%20on%20AI%20and%20other%20emerging%20technologies_Tan.pdf

141 Huawei. 5G power whitepaper. 2019. https://carrier.huawei.com/~/media/CNBG/Downloads/Spotlight/5g/5G-Power-White-Paper-en.pdf

142 Huawei. Green 5G: building a sustainable world. 2020. https://www.analysysmason.com/research/content/white-papers/green-5g-sustainability-rma18-rdns0/

143 Bjornson E, Sanguinetti L, Hoydis J, et al. Optimal design of energy-efficient multi-user MIMO systems: is massive MIMO the answer? IEEE Trans Wirel Commun, 2015, 14: 3059–3075

144 Sheth J, Bowers S M. A four-way nested digital doherty power amplifier for low-power applications. IEEE Trans Microwave Theor Techn, 2021, 69: 2782–2794

145 Niu Z S. TANGO: traffic-aware network planning and green operation. IEEE Wirel Commun, 2011, 18: 25–29

146 3GPP. RWS-210447: network energy saving and green operation for NR. 2021. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_AHs/2021_06_RAN_Rel18_WS/Docs/RWS-210447.zip

147 Mandelli S, Lieto A, Baracca P, et al. Power optimization for low interference and throughput enhancement for 5G and 6G systems. In: Proceedings of IEEE Wireless Communications and Networking Conference Workshops, 2021. 1–7

148 Na D, Choi K. Low PAPR FBMC. IEEE Trans Wirel Commun, 2018, 17: 182–193

149 El Hassan M, Crussiere M, Helard J F, et al. EVM closed-form expression for OFDM signals with tone reservation-based PAPR reduction. IEEE Trans Wirel Commun, 2020, 19: 2352–2366

150 Xu J, Zhu P, Li J, et al. Secrecy energy efficiency optimization for multi-user distributed massive MIMO systems. IEEE Trans Commun, 2020, 68: 915–929

151 Sun Y, Song C, Yu S, et al. Energy-efficient task offloading based on differential evolution in edge computing system with energy harvesting. IEEE Access, 2021, 9: 16383–16391

152 Ulukus S, Yener A, Erkip E, et al. Energy harvesting wireless communications: a review of recent advances. IEEE J Sel Areas Commun, 2015, 33: 360–381

153 Lee H S, Lee J W. Adaptive traffic management and energy cooperation in renewable-energy-powered cellular networks. IEEE Syst J, 2020, 14: 132–143

154 Huawei. 5G network architecture: a high-level perspective. 2016. https://www-file.huawei.com/-/media/corporate/pdf/mbb/5g_nework_architecture_whitepaper_en.pdf?la=en

155 NTT Docomo. White paper: 5G evolution and 6G. 2021. https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v3.0.pdf

156 Azimi S M, Simeone O, Sengupta A, et al. Online edge caching and wireless delivery in fog-aided networks with dynamic content popularity. IEEE J Sel Areas Commun, 2018, 36: 1189–1202

157 Ateniese G, Mancini L V, Spognardi A, et al. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers. Int J Netw Secur, 2015, 10: 137–150

158 Lal N, Tiwari S M, Khare D, et al. Prospects for handling 5G network security: challenges, recommendations and future directions. In: Proceedings of the 2nd International Conference on Smart and Intelligent Learning for Information Optimization, 2020. 1–8