Evolution of RAN Architectures Toward 6G: Motivation, Development, and Enabling Technologies

Jiacheng Chen, *Member, IEEE*, Xiaohu Liang, *Member, IEEE*, Jianzhe Xue, *Student Member, IEEE*, Yu Sun, *Graduate Student Member, IEEE*, Haibo Zhou, *Senior Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—In this survey paper, we first provide insights on the evolution of radio access networks (RANs) through investigating the existing paradigms and future trends towards 6G. We then present the fully-decoupled RAN (FD-RAN), which aligns with the trends by integrating existing paradigms and introducing new features such as physical decoupling of uplink and downlink base stations. We also discuss the key technologies enabled by different architectures for further performance improvement and some open issues. We hope that this survey can stimulate more in-depth research on transforming 6G RAN so as to unleash the power of state-of-the-art technologies and meet higher performance requirements in the future.**

***Index Terms*—6G RAN, fully-decoupled RAN, C-RAN, cell-free network, space-air-ground integrated network, integrated sensing and communications.**

I. Introduction

CELLULAR networks have dominated terrestrial mobile communications since the launch of the first commercial cellular communication system in 1979. The advantages, including high capacity, low energy cost of mobile devices, and large coverage, have led to the continuous evolution from 1G to the current 5G. Over the years, technological advancements such as the switch to a fully digital system in 2G, the introduction of packet switching, orthogonal frequency division multiplexing (OFDM), and multiple-input multiple-output (MIMO) in 4G, and the cloud-native deployment of core network functions in 5G [1], [2] have greatly contributed to the success of cellular networks. The increasing capabilities of cellular networks and mobile devices have also led to the prosperous development of mobile applications and services, which have become essential for our daily lives. A whole picture of 5G and beyond is given in Fig. 1, illustrating the applications, scenarios, and underlying radio access technologies.

With the worldwide rollout of 5G in the past few years, we have once again come to a critical time point where the next generation cellular network, i.e., 6G, needs to be defined. We can see that a variety of 6G research initiatives have been launched globally, such as:

- Hexa-X [3]: a European Union research project aimed at defining the research challenges and technological requirements for 6G, as well as developing a roadmap for its deployment.
- Next G Alliance [4]: a U.S.-based industry alliance formed by leading technology companies, including AT&T, Apple, Nokia, and Qualcomm, to accelerate the development and deployment of 6G wireless technologies.
- 6G Flagship [5]: a research project led by the University of Oulu in Finland that aims to create a global ecosystem for 6G research and innovation, involving academia, industry, and public authorities.

Besides, academic research also attempts to shape the future 6G, and there have been several outstanding surveys on the development of key technologies [6], [7]. Potential directions include green and sustainable 6G, metamaterials, integrated sensing and communications, non-terrestrial networks, core-edge computing integration, AI [8], digital twin [9], and blockchain [10].

A. Challenges of 6G RAN

Certainly, 6G has to confront with a variety of challenges. In the scope of radio access network (RAN), we try to focus

Manuscript received 21 August 2023; revised 26 December 2023 and 25 February 2024; accepted 28 March 2024. Date of publication 15 April 2024; date of current version 23 August 2024. This work was supported in part by the National Natural Science Foundation Original Exploration Project of China under Grant 62250004; in part by the Natural Science Foundation of China (NSFC) under Grant 62271244 and Grant 61901516; in part by the Innovation and Entrepreneurship of Jiangsu Province High-Level Talent Program; in part by the Summit of the Six Top Talents Program of Jiangsu Province; in part by the Major Key Project of PCL; in part by the National Key Research and Development Program of China under Grant 2018YFB1801103; in part by the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu Province under Grant BK20192002; and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). *(Corresponding author: Haibo Zhou.)*

Jiacheng Chen is with the Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: chenjch02@pcl.ac.cn).

Xiaohu Liang is with the School of Communication Engineering, Army Engineering University, Nanjing 210000, China, and also with the School of Information Science and Engineering, Southeast University, Nanjing 210000, China (e-mail: liangxiaohu688@163.com)

Jianzhe Xue, Yu Sun, and Haibo Zhou are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: jianzhexue@smail.nju.edu.cn; yusun@smail.nju.edu.cn; haibozhou@nju.edu.cn).

Xuemin Shen is with the Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/COMST.2024.3388511

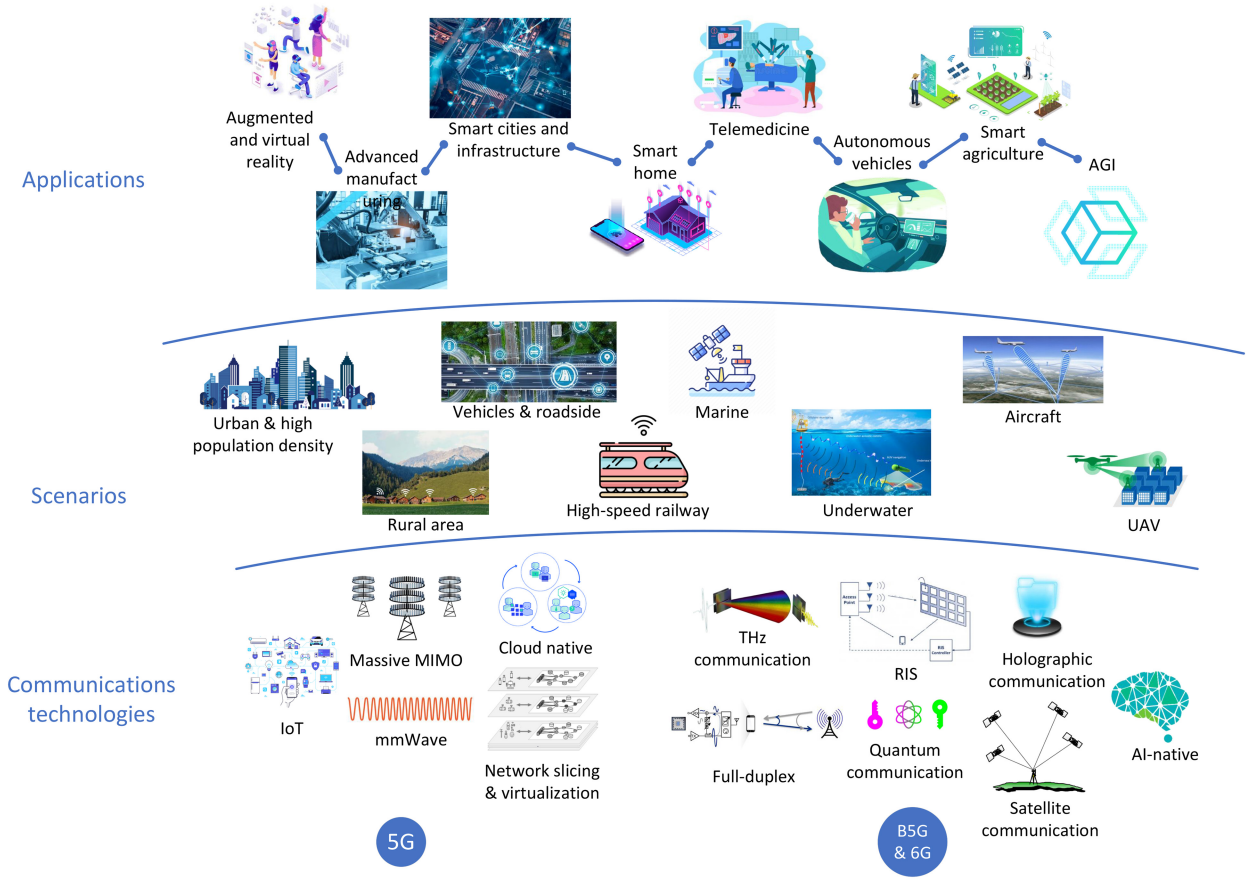

Fig. 1. A whole picture of 5G and beyond.

on the fundamental and long-standing challenges, which not only exist in 6G, but also exist in 5G and early generations. Since these challenges have not been completely addressed yet, we believe that evolution of RAN architecture is required to solve them. Meanwhile, these challenges still need to face 6G's new situation and requirements.

Firstly, in order to meet the Immersive Communication (extension of eMBB in 5G) requirement of 6G, higher data rates are necessary, which usually means using more spectrum, given that the spectrum efficiency has been highly optimized in 5G. Unfortunately, high-quality spectrum resources, especially low-frequency bands, have been exhausted. Although alternative spectrum, including millimeter waves, terahertz bands, and visible light, have been considered, their practical application scenarios are still very limited in 5G [11] due to the high frequency ranges, and it is evident that 6G cannot completely rely on these additional spectrum. Hence, it is worth further improving the utilization efficiency of the existing low-frequency bands available for the cellular networks. However, the inflexibility in the current spectrum utilization pattern often leads to spectrum shortage (i.e., demand is higher than supply) in both time and space domains [12]. Specifically, in both frequency division duplex (FDD) and time division duplex (TDD) modes, the same spectrum can only be used for either uplink or downlink at a given time, and the need to guard between uplink and downlink adds complexity and consumes additional resources. Additionally, operators are restricted to using only their licensed spectrum, leaving their spectrum supply fixed, while the traffic demand from their subscribers can experience significant spatiotemporal fluctuations.

Another critical issue is the heavy cost burden associated with deploying and operating mobile networks, including capital expenditures (CapEx) and operating expenditures (OpEx). This challenge becomes ever dominant for mobile operators towards the 6G era, because 5G has introduced much higher costs than the previous generations. The increase in CapEx is primarily due to the cost of more powerful hardware, such as base stations (BSs) equipped with large antenna arrays, the cost of new licensed spectrum, and the cost of deploying additional BSs for ultra-dense networks. The increase in OpEx mainly stems from the higher energy consumption required for data transmission and processing. It has been reported that the power required by a single 5G BS is approximately four times that of a 4G BS [13], and the higher density of BSs exacerbates the energy cost burden. Since 6G is anticipating more high-cost extensions from 5G and introducing new features, such as extra-large scale massive MIMO [14], [15], reconfigurable intelligent surface (RIS) [16], and AI-related applications [17], it is imperative to figure out how to achieve the balance

between the investment and potential profits of 6G, so as to deploy 6G in a more cost-efficient way [18] that is acceptable for both mobile operators and users.

Finally, 6G is expected to be truly everyone-centric [19] and provide customized service to users. Although garnered very little attention, the current 5G mobile networks still operate with the primary focus on improving transmission performance, while totally neglecting users' actual needs and desires. In 5G, the network is not aware of the diverse and varying service requirements of the users, and there is basically no mechanism to guarantee a specific user's quality-of-experience (QoE). Thus, more stringent requirements for 6G, e.g., hyper reliable and low-latency communication (HRLLC, extension of uRLLC in 5G), cannot be satisfied. It is important to recognize that the ultimate goal of the network is to meet users' needs for information and services, and bring more value to them, rather than simply transmitting larger amounts of data without considering their value to the users. Thus, the future cellular network should shift to a user-centric paradigm [20] and offer personalized services.

B. Motivations

To address these long-standing challenges, we believe that the underlying RAN architecture of 6G needs to be transformed. 5G RAN still inherits from its early generations, utilizing cell as the basic unit for composing large-scale networks. User control and uplink/downlink data transmission are restricted in each individual cell. Further, user's serving cell changes through handover. The single cell based architecture is simple and easy for RAN deployment and user management. However, it also limits the potential for more advanced physical and MAC layer technologies.

The motivation of this survey is to emphasize the necessity of transforming the underlying architecture of 6G RAN and inspire more in-depth thoughts and discussions on this topic. This is critical especially at this early stage of 6G standardization and development. To find out how the 6G RAN architecture should be transformed, we present this survey from the perspective of RAN architecture evolution. We first investigate the existing paradigms that have been widely acknowledged for their advantages. In particular, our focus will be on the centralized control paradigm, heterogeneous cell paradigm, and no cell paradigm. The RANs corresponding to these paradigms are known as cloud RAN (C-RAN), heterogeneous network (HetNet), and cell-free (CF) network, respectively. Then, we derive the trends of 6G RAN evolution from the existing paradigms, considering the long-standing challenges. Finally, we present a tutorial on a newly-proposed fully-decoupled RAN (FD-RAN) architecture [35] that aligns with these trends by adopting the decoupling paradigm. FD-RAN unifies the above RANs and is a promising direction of 6G RAN evolution. In the following, we briefly introduce the architecture and motivations of these RANs, and in particular, highlighting the differences and advantages of FD-RAN.

1) Centralized Control Paradigm: In C-RAN, signal processing is separated from signal transmission and centralized. The baseband processing units (BBUs) handle signal processing while remote radio heads (RRHs) are responsible for signal transmission. BBUs are centralized and run on the edge cloud to reduce operational costs, while RRHs are distributed at various locations to ensure that users are covered by signals. BBUs and RRHs are connected through high-speed fiber optic fronthaul links. More generally, since data processing is centralized, it is natural to control the wireless resource allocation of these RRHs to further reduce interference.

2) Heterogeneous Cell Paradigm: In HetNet, cells of varying sizes are deployed and their coverage areas may overlap. These cells can be categorized into macro cells, small cells, femto cells, etc. based on their size. These cells have different characteristics and are suited for different application scenarios. For example, small cells are mainly used for indoor environments and macro cells for outdoor environments. More generally, the cells can be heterogeneous in terms of their used spectrum bands, e.g., sub-6GHz for macro cells and mmWave for small cells. Also, the cells can be non-terrestrial as in aerial radio access network (ARAN) comprising airborne objects employed for transmission [33].

3) No Cell Paradigm: The cell-free network is similar to a distributed massive MIMO system and follows the C-RAN paradigm. Different from C-RAN, many lightweight access points (APs) are uniformly distributed and controlled by central processing units (CPUs). Each user is served by multiple APs simultaneously, with the MIMO transmission technology providing support. This ensures that all users receive the same level of quality-of-service (QoS) and there is no cell edge as in traditional cellular networks.

4) Decoupling Paradigm: Inspired by neurotransmission [36], in FD-RAN, a conventional fully functional base station is split into three different types of base stations: downlink data base station (DL-BS), uplink data base station (UL-BS), and control base station (C-BS). The DL-BSs and UL-BSs are deployed separately, based on the needs for data reception or transmission, while the C-BS covers a larger area and exchanges low-latency, low-volume control-related data with users. For a user equipment (UE), control messages and data transmission are separately transmitted on different physical channels. Furthermore, uplink and downlink can be served by different base stations, and cooperative transmission techniques in both directions can be applied.

In the decoupling paradigm, control and data transmission are physically separated, and also uplink and downlink data transmission are physically separated. Thus, FD-RAN enhances the centralized control of C-RAN with the dedicated control BS and typically a larger control area. FD-RAN also brings more flexibility to efficiently integrate the heterogeneous resources in HetNet owing to the complete decoupling of uplink and downlink networks. FD-RAN can further reduce the deployment costs of CF networks since UL-BSs and DL-BSs are no longer required to be co-located. Compared with traditional RAN, spectrum can be utilized for either uplink or downlink in FD-RAN, and users' personalized service requirement can be satisfied through resource cooperation. Despite being a new architecture, FD-RAN has attracted attention from both academia [6] and industry [37]. Also,

TABLE I
SUMMARY OF SELECTED RAN SURVEYS

Ref.	RAN Generation	Main Discussed Topics	Differences on Perspectives, Insights or Proposals
[21]	5G	New radio access technologies, emerging system-level technologies, and interplay between them.	Proposes that these technologies should complement so as to enable versatile and adaptable 5G networks.
[22]	5G	A detailed description of each functional split option and assessment of advantages and disadvantages.	Provides insights on how the fronthaul network will be affected by the choice of functional split.
[23]	6G	General technologies and challenges for 5G and 6G.	Proposes a virtualized network slicing based architecture of 6G.
[6]	6G	In-depth survey of 6G, including vision, requirements, architectures, key technologies, testbeds, etc.	Provides insights on lessons learned from literature.
[7]	6G	Fundamental 6G technologies for IoT. 6G for IoT applications.	Provides insights on the convergence of 6G and IoT.
[24], [25]	5G	Surveys on architecture and technologies of C-RAN.	Focuses on the C-RAN perspective.
[26], [27]	5G	Survey on resource allocation in HetNets for 5G communications.	Provides two potential structures to solve the RA problems of the next-generation HetNets.
[28]	5G	General survey on 5G and its technologies. Discussion on various C-RAN like architectures.	Provides comparisons on different architectures from various perspectives.
[29]	5G	General survey on ultra-dense networks (UDN).	Focuses on the UDN perspective.
[30]	5G	A holistic overview on the DL–UL decoupled access (DUDe) approach.	Focuses on the DUDe perspective.
[31]	5G	SDN application in the context of 5G communication and C-RAN.	Focuses on the SDN-C-RAN integration perspective.
[32]	5G	DL techniques applied in C-RAN.	Focuses on the convergence of C-RAN and DL.
[17]	6G	The cloud-edge-mobile infrastructure and technologies required to support AIGC services.	Provides insights on edge intelligence and mobile AIGC.
[33]	6G	Survey on ARAN architecture, features and technologies.	Focuses on the ARAN perspective.
[34]	5G/6G	Comprehensive survey on O-RAN from various aspects.	Provides insights on the fundamental logic of O-RAN.
Ours.	6G	Existing paradigms and trends of RAN. Architecture-enabled technologies. A newly-proposed FD-RAN architecture for 6G.	Provides a unified RAN evolution perspective. Provides insights on existing paradigms and future trends. Proposes to transform the underlying RAN architecture of 6G to enable more technologies.

the potential application scenarios of FD-RAN cover some important emerging services for 6G, such as space-air-ground integrated network, integrated sensing and communications, and AI generated contents.

C. Our Perspective and Differences With Related Surveys

The concept of RAN architecture discussed in this survey is from the high-level networking design perspective. We investigate different modalities of organizing BSs into large-scale RANs, and focus on their motivations and enabled technologies. On the other hand, the RAN architecture mentioned in the context of standardization bodies such as 3GPP or O-RAN Alliance mainly refers to the specifications of functional splits [22] at hardware or software level. Particularly, O-RAN will evolve RAN from different perspectives [34]. Similar to 3GPP 5G NR, it disaggregates BSs into functional units and defines corresponding interfaces, yet aiming at improving the interoperability and hardware flexibility, such that multiple vendors can participate in the market to lower the costs. It also defines the RAN intelligent controllers (RICs) to facilitate programmability and AI/ML.

Since existing surveys have discussed various topics of RAN, we present a brief review on related surveys, especially with unique perspectives, insights or proposals, and highlight our contributions and differences with them through Table I. *To summarize, our survey offers a unified study of different RAN paradigms and their enabled technologies, highlighting the key point that an appropriate architecture is the foundation for enabling new technologies. The point is further demonstrated through the decoupling paradigm and the FD-RAN architecture, which not only integrates the existing paradigms but also develops its own features, pushing forward the RAN evolution to achieve technology deployment and versatility.* The main contributions of the survey are highlighted below:

- We discuss the evolution of RAN architectures towards 6G by investigating the existing paradigms and point out the future trends. We further introduce the architecture of FD-RAN and show how it is aligned with the trends.
- We present surveys on the key technologies that are appropriate to be deployed in 6G so as to meet the future performance demands. We show how these technologies are enabled by the underlying RAN architectures, as well as the lessons learned from these technologies.
- We present case studies on the key technologies under the FD-RAN architecture, and show how FD-RAN enables these technologies through the lessons learned. We also discuss the future directions of FD-RAN, including the integration of FD-RAN with other emerging services in 6G, and open issues in FD-RAN.

D. Organization

The organization of the survey is outlined in Fig. 2. In the remainder of this survey, we first examine the RAN paradigms. An overview of the existing RAN paradigms are presented in Section II. The trends of RAN evolution are summarized in Section III, together with the newly-proposed decoupling

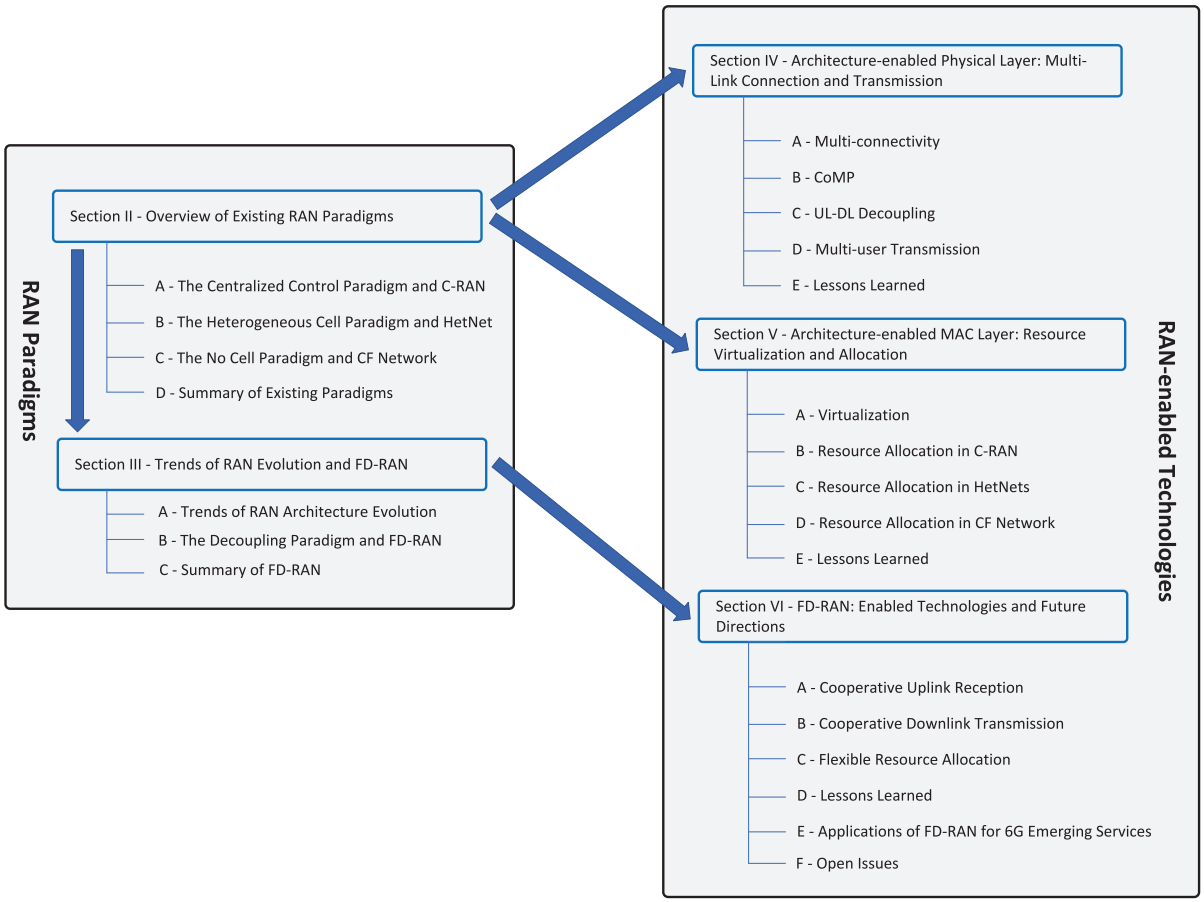

Fig. 2. Organization of the survey.

paradigm and FD-RAN architecture that integrates the existing RANs. Then, architecture enabled PHY- and MAC-Layer technologies are discussed in Sections IV and V for existing RANs, and in Section VI for FD-RAN, respectively. We draw lessons from these three sections, and present future applications and open issues for FD-RAN in Section VI. Conclusion of the entire survey is given in Section VII.

II. Overview of Existing RAN Paradigms

This section presents a brief review of the existing RAN paradigms with representative features.

A. The Centralized Control Paradigm and C-RAN

One of the practical problems for wireless network deployment is the lack of equipment rooms, as air conditioning is essential for BS deployment. Besides, in order to fulfill increasing user demand, more BSs will have to be deployed in the same region, which may result in high inter-cell interference if these BSs are not coordinated. The densely deployed BSs will also lead to high energy consumption and high maintenance cost.

To overcome the above challenges, C-RAN was conceived to obtain flexible network capabilities while curtailing network deployment and maintenance expenditures. The concept was first proposed by IBM in 2010 with the name of wireless network cloud (WNC), building on the concept of Distributed Wireless Communication System [38]. Further, its technology trends and feasibility analysis is elaborated in 2011 by China Mobile Research Institute [39].

In the early days of cellular networks, radio frequency and baseband processing functionalities were integrated into a single BS in 1G and 2G networks. As technology advanced, the BS was decoupled into RRHs and BBUs in 3G and 4G networks. This allowed for lower site rental and ease of maintenance. In beyond 4G and in 5G, with the emergence of cloud computing and data centers, BBUs are virtualized into BBU pools that are shared by multiple base stations. RRHs are connected to the BBU pools via fronthaul links. This evolutionary process has led to the development of a common C-RAN architecture that leverages the benefits of centralized processing and virtualization to improve network performance and efficiency.

The C-RAN architecture is shown in Fig. 3, and it consists of four main components: RRH, fronthaul, BBU pool, and backhaul. RRH and BBU pool are functional entities, while fronthaul and backhaul represent the two connection tiers. Below, we provide a brief overview of each component:

- *RRH:* The RRH component is responsible for transmitting radio frequency signals to users, which involves tasks such as filtering, RF amplification, and A/D and D/A conversion. RRHs serve as a simple RF front-end, with the majority of processing tasks performed in the BBU pool. This design makes RRHs easy to

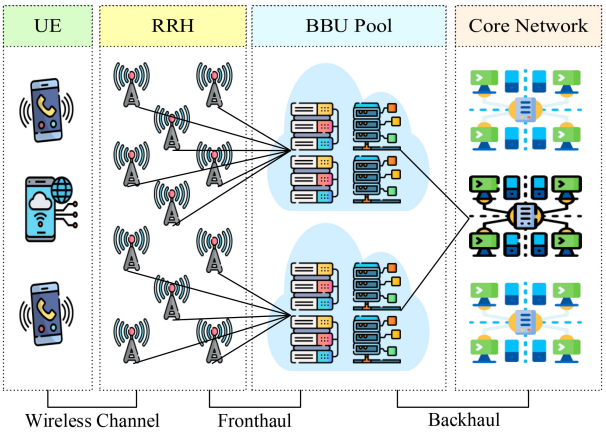

Fig. 3. C-RAN system.

deploy and allows for extensive distribution at a low cost.

- *Frouthaul:* The fronthaul is the connection tier that links a set of RRHs to the BBU pool, providing high-bandwidth links to handle the requirements of multiple RRHs. Fronthaul communication technologies include optical fiber communication, millimeter-wave communication, and cellular communication. Among them, optical fiber is considered the most ideal communication approach in C-RAN as it can provide the highest bandwidth. The RRHs communicate with the BBU pool via the Common Public Radio Interface (CPRI) protocol, which is a radio interface protocol widely used for transmitting in-phase and quadrature (IQ) data between RRHs and BBUs.
- *BBU pool:* The BBU pool is a collection of virtualized computing units that can be located at a centralized site, such as a cloud server or data center. By performing baseband processing in a centralized manner, it can facilitate cooperative radio resource allocation, collaborative processing at large-scale, and intelligent networking.
- *Backhaul:* The backhaul is responsible for transmitting data between the BBU pool and the mobile core network. This can include both wired and wireless communication technologies, depending on the network infrastructure and the specific deployment scenario. The backhaul is an important component of the C-RAN architecture, as it enables the BBU pool to communicate with the core network and provide services to end users.

The C-RAN architecture introduces significant demands on the fronthaul transport network, which must support high bandwidth, low latency, and low jitter requirements. In light of the exacting demands imposed by the C-RAN architecture, several system structures have been explored to alleviate the strain on the fronthaul transport network by transferring certain processing tasks from the BBU pool to the RRHs. In addition, partial centralization has emerged as a feasible trade-off to offset the costly initial investment required to transition from conventional distributed networks to C-RAN, while simultaneously curtailing operational expenses by aggregating baseband processing to the cloud and minimizing power consumption [40]. As shown in Fig. 4, the basic difference between

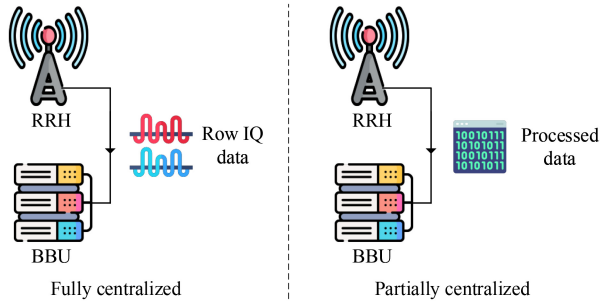

Fig. 4. Fully and partially centralized architecture.

these C-RAN system structures lies in the location where the physical layer functions are handled, resulting in what kind of signals are transmitted in fronthaul. The C-RAN system structures can be classified into three categories, namely fully centralized, partially centralized, and hybrid, based on the functionalities of the BBU pool and RRH:

- *Fully centralized architecture:* In a fully centralized C-RAN architecture, all processing tasks related to the physical layer are handled entirely by the BBU pool located in a centralized data center. The fronthaul carries raw digitized radio signals between the RRHs and the BBU pool. This architecture is advantageous in terms of ease of operation and maintenance since all the functions of managing and processing resources are centralized at the BBU pool. However, it requires significant bandwidth and low latency on the fronthaul links to accommodate the transfer of raw data between the RRHs and the BBU pool.
- *Partially centralized architecture:* In a partially centralized C-RAN structure, some physical layer processing is done at the RRH to alleviate the burden on the fronthaul, as the physical layer information accounts for a significant portion of communication resources. For instance, the RRH only receives the demodulated signals, which require only 20–50 times less bandwidth than the modulated signals. However, such a structure may not deliver optimal performance due to reduced resource sharing and the inability to effectively support advanced features such as cell-free.
- *Hybrid architecture:* The hybrid C-RAN architecture seeks a middle ground between the fully centralized and partially centralized structures. It offers the ability to shift processing loads dynamically between the RRHs and BBUs to optimize for fronthaul bandwidth usage, latency, and network conditions. It aims to maximize the network's operational efficiency and support for advanced features while managing the fronthaul transport requirements more effectively than the fully centralized model.

C-RAN offers numerous advantages in various aspects. The centralized BBU pool brings several benefits, including simple equipment for RRHs, efficient utilization of BBUs, cost-effective network deployment and operation, and ease of updates. Additionally, C-RAN provides flexibility for network scalability and adaptability to non-uniform traffic patterns.

Furthermore, centralized baseband processing can improve energy efficiency and spectrum efficiency by introducing advanced technologies such as interference mitigation and CoMP. These advantages are discussed in detail below.

- *Ease for network upgrades and maintenance:* The rapid growth of mobile data requires the mobile operators to significantly increase communication capacity of their networks in order to offer mobile broadband to the general public. As a results, the total cost of ownership (TCO) is increased, which includes OpEx and CapEx [24], [39]. By centralizing compute tasks in a few data centers and offloading easy tasks to RRHs, C-RAN reduces operating and maintenance costs. The RRH in C-RAN is much simpler and cheaper than the BS in traditional RAN, requiring less hardware for installation and no cooling system for operation. Additionally, C-RAN enables more effective equipment sharing, resulting in lower CapEx. Moreover, centralized BBU pools allow for more regular hardware updates than BBUs located in remote locations, potentially benefiting from IT advances in CPU technology. Quantitative cost analysis shows that C-RANs can reduce capital expenditure per kilometer by 10% to 15% compared to traditional LTE networks [41].
- *Agile network architecture:* The rapid growth of mobile communication demands requires mobile networks to dynamically update their coverage and capacity. C-RAN presents a seamless approach for network coverage extension and technology evolution. Its centralized structure simplifies scaling of cellular network coverage. Network operators can connect new RRHs with BBU pools to expand service regions and add new compute units into BBU pools to enhance baseband processing ability. Additionally, a more decoupled structure facilitates easier and cost-effective network hardware equipment and communication standards updates. The BBUs in the C-RAN is centralized in a few locations, making it easier to update than if they were placed in a decentralized manner. Meanwhile the RRHs is only has the function of radio frequency and it is cheap to update. Additionally, C-RAN can enhance network capacity by dividing cells into smaller cells, which can improve coverage and increase capacity for high-traffic areas. Besides, the software updates, such as resource allocation algorithms and communication standards, can be quickly achieved in BBU pools.
- *Adaptability to nonuniform traffic:* The daily traffic distribution in each cell is dynamic, and traffic peaks occur at different times. Traditionally, the maximum BBU capacity of base stations often depends on the peak traffic hours, causing a huge amount of wasted when users move from busy areas to residential areas [42]. In C-RAN, the BBU pool is responsible for a large region, allowing for optimal allocation of compute resources to RRHs based on their instantaneous demand. Despite the dynamically changing serving RRH due to UE movement, the serving BBU can remain in the same BBU pool. Non-uniformly distributed traffic caused by movement can be handled in the same BBU pool, as it contains a set of RRHs with much larger coverage compared to traditional BSs. This allows for high computing resource efficiency of the BBUs, even in the presence of non-uniform traffic. An evaluation of statistical multiplexing gains indicates that the number of BBUs in the Tokyo metropolitan region can be reduced by up to 75% compared to a typical RAN architecture [43]. Another study found that by leveraging the variation in processing demand between base stations, using centralized processing can save more than 22% of computing resources [44].
- *High spectrum efficiency:* C-RAN with centralized control has enabled many new technologies to improve the spectrum efficiency, which cannot be achieved in the traditional RAN with distributed BBUs. In the traditional LTE systems, all cells operate at the same frequency, which can result in considerable inter-cell interference. For instance, the observed ratio of peak to cell edge throughput can reach up to 10 [24]. In C-RAN, signal processing of many neighboring cells can be handled in the same BBU pool, easing the implementation and reducing processing and transmitting delays. With the BBU pool, communication information such as channel state information (CSI) and traffic data of UEs can be easily shared among RRHs with low latency. Therefore, joint processing and scheduling can be implemented, improving the spectral efficiency of the network by minimizing inter-cell interference and optimizing resource allocation.
- *High energy efficiency:* Energy consumption in mobile networks is primarily used for powering amplifiers, RRHs, BBUs, and air conditioning. Deploying C-RAN can improve the energy efficiency of the mobile network in several ways. Firstly, the C-RAN architecture can reduce the number of BS sites, resulting in lower power consumption for air conditioning and other supporting equipment. Secondly, since traditional networks consume most of their energy on BSs, moving the functions of BSs to a centralized cloud server can improve the utilization rate of processing resources and reduce power consumption [45]. In addition, C-RAN architecture offers the opportunity to selectively turn off machines in the BBU pool during periods of reduced processing demands, such as at night. Furthermore, by utilizing higher density RRH networks and cooperative radio management techniques, interference between neighboring RRHs can be minimized. In total, comparing with traditional RAN architecture, ZTE has claimed that C-RAN can reduce 67%–80% power consumption while China Mobile research estimates 71% for power savings [39], [46].

Regarding standardization, the concept of C-RAN was first introduced in LTE and its architecture is defined in the Technical Specifications (TS) 36.201 and TS 36.300 series. These specifications outline the functional split between the BBU and RRH, interoperability requirements, and protocols for communication between different C-RAN components. Further, the C-RAN architecture in 5G NR is defined in the TS 38.401 and TS 38.300 series. These specifications provide

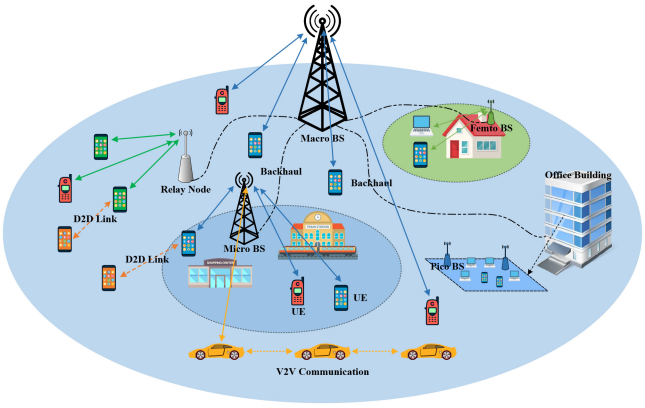

Fig. 5. The architecture of future HetNets.

guidelines for the functional split options, fronthaul interfaces, and the interaction between central units (CUs) and distributed units (DUs).

B. The Heterogeneous Cell Paradigm and HetNet

The burgeoning data demand in next-generation networks has prompted the exploration of dense cell deployment as a promising solution. However, increasing the number of identical macro cells directly would result in higher interference and lower service quality. Instead, deploying smaller and simpler BS to complement macro cell coverage is a more suitable alternative, resulting in a heterogeneous network with various cell types, which is in contrast to homogeneous networks consisting of identical cells. Furthermore, recent advancements in communication technologies, such as machine-to-machine (M2M) communication, Internet of Things, and device-to-device (D2D) communication, have led to increasingly diverse services and personalized communication requirements, further amplifying network heterogeneity.

Traditional HetNets typically incorporate macro BSs (MBSs) and small BSs (SBSs). However, with the emergence of diverse communication technologies, the architecture of future HetNets is becoming more heterogeneous, as illustrated in Fig. 5. Specifically, we classify these communications into the following three categories.

- MBS communications refer to traditional communication methods that rely on legacy BSs (e.g., 2G/3G BSs), which constitute the foundation of communication in HetNets. MBSs are responsible for providing wide coverage and seamless mobility for UEs using high transmit power over large geographic areas. The maximum radius of coverage area varies from 1 km to 25 km, depending on the BS type and terrain.
- The introduction of SBS communications on the basis of MBS communications marks the beginning of the heterogeneous phase of cellular networks. SBSs are deployed within the coverage area of MBSs to provide expansive coverage, increased capacity and higher-speed transmission. They also help to mitigate the traffic burden of MBSs, which have lower transmit power and smaller coverage areas than MBSs. SBSs encompass various types of BSs, including micro BSs, pico BSs, and femto BSs, each with different coverage, power, number of served users, and scenarios. A detailed comparison of MBSs and various SBSs is shown in Table II.
- While traditional HetNets only incorporate MBSs and SBSs, as stated in 3GPP Release 12 [47], there is a growing range of diverse communication technologies that could also be integrated into HetNets to embrace a wide range of applications. These emerging technologies comprise relay communication, D2D communication [48], vehicle-to-vehicle (V2V) communication [49], and more. In addition, there has been discussion of combining multiple radio access technologies (RATs) such as WiFi and Bluetooth [50], as well as multiple bands such as mmWave. These emerging communications represent further heterogeneity in future HetNets.

TABLE II
COMPARISONS OF MBS AND SBSS [26]

BS Types	Coverage	Power (W)	Served Users	Scenarios
Macro	few 10s kilometers	20 to 160	100s	urban, rural
Micro	few 100s of meters	5 to 20	>128	shopping malls, railway station
Pico	10s of meters	0.1 to 5	64 to 128	office building
Femto	10s of meters	0.01 to 1	4 to 32	home, small office

With the proliferation of heterogeneous cells, the capacity of HetNets will be significantly increased. SBSs further improve the service quality of edge users, filling coverage holes and expanding coverage areas. SBSs also shorten the distances between users and BSs, leading to reduced pass loss, delays, and power consumption in HetNets. HetNets enable MBSs to share spectrum with SBSs, which tremendously improves spectral efficiency. Emerging techniques, such as D2D and V2V communications, can further facilitate the efficiency by utilizing spectrum sharing [26]. In terms of operational costs, HetNets have been found to be more cost-effective compared to homogeneous networks with the same density. This is attributed to the higher efficiency of deploying and operating SBSs in HetNets. Moreover, the proper offloading of traffic to SBSs has been shown to diminish congestion in MBSs, thereby potentially improving overall system performance [51]. HetNets have demonstrated a multitude of advantages, which will be further amplified and augmented through the enabling technologies.

In order to address the severe interference issues present in HetNets, as well as the limitations of CoMP transmission, a new solution known as heterogeneous cloud radio access networks (H-CRAN) has been proposed [52]. This innovative approach seeks to leverage the benefits of C-RAN, such as the powerful and efficient cloud computing capabilities, to enhance cooperative gains and reduce interference. Fig. 6 illustrates the architecture of H-CRAN, where MBSs are classified as high-power nodes (HPNs), and SBSs and RRHs are categorized as low-power nodes (LPNs). H-CRAN leverages control-data separation techniques, with HPNs providing the control plane function and enhanced coverage, and LPNs

TABLE III
COMPARISONS OF C-RAN AND H-CRAN

Comparisons	C-RAN	H-CRAN
Architecture	Centralized with homogeneous distributed units, only including RRHs	Centralized with heterogeneous distributed units, including HPNs (MBSs) and LPNs (SBSs and RRHs)
Control-Data Separation	No, control and data plane are coupled	Yes, HPNs provide the control plane function and LPNs are responsible for the data plane
Coordination	One-level coordination, only between RRHs	Multi-level coordination, between HPNs and LPNs
Deployment	Less flexible and less scalable, restricted by homogeneous RRHs and coupled control and data planes	More flexible and more scalable, heterogeneous distributed units and separated control-data plane enable high flexibility and scalability
Fronthaul Requirements	Strict, low-latency & high capacity	Alleviated requirements by separated control-data plane
Energy Efficiency	Good, benefit from centralized operation	Better, not only from centralized operation but also heterogeneous deployment and sleeping of LPNs

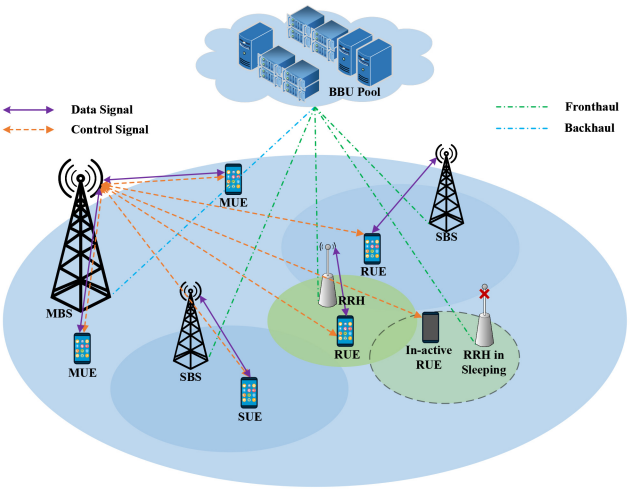

Fig. 6. The architecture of H-CRAN.

responsible for the data plane, diverse communication requirements, and improved system capacity. The centralized BBU pool is used to coordinate HPNs and LPNs, process data, and manage interference efficiently. HPNs and LPNs are connected to the BBU pool via high-speed, low-latency fronthaul links, such as optical fiber. It is worth noting that HPNs also connect to the BBU pool to coordinate with LPNs and mitigate interference.

H-CRAN presents numerous advantages over both HetNets and C-RAN, making it an attractive option for network deployment. One of its key strengths is the ability to enhance system capacity and coverage by flexibly deploying additional LPNs as required [28], an approach that is quite complicated in C-RAN due to its lack of a decoupled data plane and HetNets' characteristics. However, in H-CRAN, operators only need to add some new LPNs and connect them to the BBU pool. Moreover, the strict fronthaul requirements that are a drawback of C-RAN are alleviated in H-CRAN. By decoupling the control plane from the BBU pool and enabling HPNs to deliver signaling to UEs, H-CRAN reduces the control latency restrictions and network capacity on fronthaul, particularly for latency requirements [53].

As a crucial aspect of RANs, we address the issue of energy consumption in HetNets as follows. The introduction of additional SBSs results in increased energy consumption. However, SBSs exhibit a higher energy efficiency due to their reduced cooling requirements. Moreover, the overall energy efficiency of HetNets surpasses that of homogeneous networks with the same density, as highlighted in [27]. This superiority can be attributed to the more flexible deployment and operation characteristics of HetNets. In addition, various technologies emerging in HetNets, including carrier aggregation [54], dual connectivity [55], CoMP [56], UL-DL decoupling [57], ICIC, and enhanced ICIC (eICIC) [58], contribute to higher energy efficiency for HetNets and have demonstrated their superiority. Regarding H-CRAN, it can leverage the energy efficiency benefits of both HetNet (e.g., flexible deployment) and C-RAN (e.g., efficient cooling systems in the BBU pool), thereby advantaging higher energy efficiency compared to C-RAN and HetNets. Furthermore, the BBU pool's management enables underutilized LPNs to sleep [59], promoting energy efficiency, a feature that is not feasible in C-RAN due to homogeneous RRHs and coupled control-data plane. H-CRAN operates similarly to C-RAN in a centralized manner, yet there are notable differences between them, as outlined in Table III.

3GPP has actively participated in the standardization of HetNets. Although information about architectures is limited, some definitions concerning scenarios and key technologies have been established. In Release 12, HetNet scenarios were defined, integrating MBSs and SBSs [47]. Numerous key technologies enabling HetNets within 3GPP have also been identified. The concept of dual connectivity was introduced by 3GPP in Release 12 [60]. With the evolution from LTE to 5G, a variety of options for multi-radio dual connectivity have been standardized in 3GPP [61]. Release 14 marked the inception of intra-site carrier aggregation [62]. Addressing a crucial aspect, the standardization of UL-DL decoupling was undertaken in Release 12 [60]. In 3GPP Release 8/9, Inter-Cell Interference Coordination (ICIC) was introduced in the standardization process, and subsequently, enhanced ICIC (eICIC) was introduced in Release 10 to further augment its capabilities [63].

C. The No Cell Paradigm and CF Network

In traditional cellular systems, the coverage area of each antenna unit is usually fixed. To reduce interference between

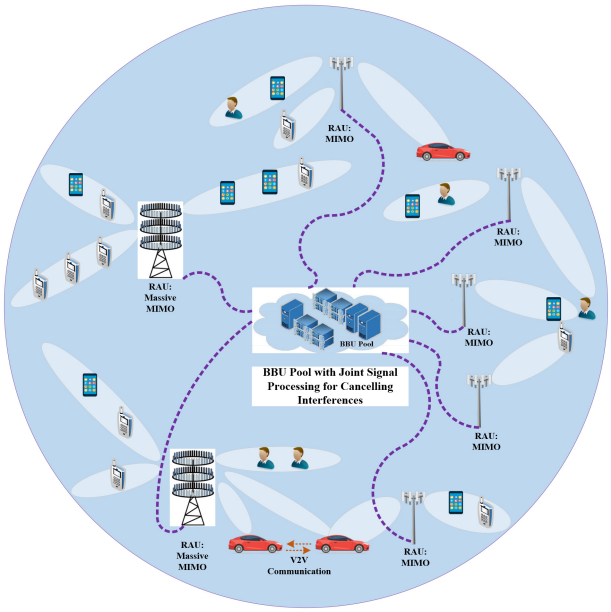

Fig. 7. The architecture of cell-free network.

cells, different frequency bands are usually allocated to different cells. Moreover, the base station does not have the ability of joint signal processing. With the increasing of cell density and the number of users in the future 6G era, the main disadvantage of traditional cellular systems is that the interference among cells will increase dramatically. Also, the system throughput will be severely limited.

Compared with the traditional cellular communication system, the main feature of the cell-free network system is that the Remote Antenna Unit (RAU) and the Base Band Unit (BBU) are separated, and the two are generally connected by optical fiber, as shown in Fig. 7. Another feature of the cell-free network system is the joint signal processing at BBU pool. Cell-free network belongs to the distributed antenna system (DAS), but with a much denser and uniform AP deployment. The applications of massive MIMO technology in CF network can further enhance system capacity and service performance. The terminology of CF massive MIMO (mMIMO) was proposed in [64], where a very large number of distributed access points simultaneously serve a much smaller number of users.

As shown in Fig. 7, in order to realize the joint signal processing, all RAUs and BBUs generally need to be connected through optical fibers. The development of cell-free network has led to the emergence of front-haul network, which is usually consist of specially designed fiber optic Ethernet. To avoid the direct connection between each RAU and BBU, all RAUs and BBUs are connected to the front-haul network. In addition, to effectively improve the performance of the system, the front-haul network needs to be carefully designed in many aspects, such as real-time, congestion control, and time synchronization.

A cell-free network deploys a large number of distributed access points that cooperatively serve users, leading to increased energy consumption, primarily due to the larger number of active nodes and the enhanced processing requirements. To mitigate this, several solutions are being explored. For example, the implementation of energy-efficient hardware and the optimization of power amplifiers can significantly reduce the energy requirement per node. Also, advanced algorithms for power control and resource allocation can optimize the network's overall energy usage. These algorithms intelligently adjust the power levels and switch off redundant nodes when not needed. The adoption of machine learning techniques for predictive maintenance and energy management can further optimize energy usage by anticipating network loads and adjusting resources accordingly.

The key advantage of cell-free network is to allow each RAU using the same time-frequency resource to provide services for multiple users [65]. The signals of each RAU are converged to the BBU through optical fiber. To eliminate the interference among RAUs, the BBU will jointly process the signals from different RAUs. According to the demands and locations of different users, it can flexibly carry out beamforming and resource scheduling. By joint signal processing, the interference among different users can be eliminated as much as possible. In this way, the cell-free network can improve spectrum efficiency, enhance system coverage, and provide uniform data rates.

To better eliminate the interference among users on the same time-frequency resource, accurate CSI acquisition is indispensable. Thus, CF network usually applies the time division duplexing (TDD) mode, where the same frequency band is applied to both the uplink and downlink and a good symmetry between the downlink and uplink channel matrix can be obtained. The specific process is that the user sends uplink detection reference signal (SRS) to the base station, and the base station receives the signal and performs channel estimation to obtain the uplink channel matrix. Then, the downlink channel matrix is obtained by using the reciprocity of the uplink channel matrix and downlink channel matrix.

In the downlink data transmission, the data need to be precoded before sending to multiple users. The common multi-user precoding methods are block diagonalization multi-user precoding algorithm, and zero-forcing multi-user precoding algorithm. The main idea of block diagonalization precoding is to map the precoding matrix design of the target user to the null space of the channel matrix of other users. The main idea of zero-forcing precoding is firstly to obtain the pseudo-inverse matrix of the whole downlink channel matrix. Then, by using the property of zero-forcing matrix, the matrix can be decomposed. In this way, we can obtain the multi-user precoding matrix of the downlink. After precoding and beamforming, the signal will be transmitted to different users on the same time-frequency resource grid. In order to maximize system capacity, it is also necessary to carry out downlink power control, considering the user's location and quality of service (QoS), the size of the coverage area, interference suppression and other factors.

In the uplink of cell-free network, different users send signals to the base station on the same time-frequency resource grid. For improving the uplink transmission performance, uplink power control method is usually applied to suppress

TABLE IV
COMPARISONS AMONG DIFFERENT RAN ARCHITECTURES

RAN architecture	BS association pattern	Spectrum and interference management
Traditional RAN: single-cell	UE is associated with a single BS	Adjacent cells use different spectrum, or there will be strong inter-cell interference
C-RAN: cell cloud with centralized control	UE can be served by one or more BSs	BSs can use the same spectrum; interference is handled by cooperative transmission techniques
HetNet: multi-cell with heterogeneity	UE's UL and DL can be decoupled and associated with separate BSs	MBSs and SBSs can use the same spectrum (underlay) or different spectrum (overlay); inter-cell interference between MBS and SBS is handled through power control
Cell-free network: no-cell and ultra-dense deployment	UE is served by multiple cooperative APs by default	Spectrum is reused among APs and massive MIMO techniques are used, so there will be no interference if the network is not overloaded
Fully-decoupled RAN: UL: no-cell DL: flexible multi-cell Control: single-cell	UE is served by at least one C-BS, and possibly multiple cooperative DL-BSs and UL-BSs	C-BS uses dedicated spectrum; UL-BSs and DL-BSs use whole spectrum band without predefined constraints; interference is handled through fine-grained resource allocation on multiple dimensions based on utilization of cooperative transmission techniques

the interference among different users. The common uplink control method is called maximum-minimum power control method, which aims to guarantee that different users can obtain better quality of service. When the AP receives signals from different users, it needs to conduct joint signal processing. The common method of joint signal processing is to convert multi-user composite channel into multi-channel parallel single-user channel, so that the interference between multi-users can be decoupled. Further, the iterative interference cancellation algorithm will be applied to reduce the multi-user interference for improving the transmission performance of the whole system. Moreover, to avoid the inversion operation of large dimensional matrix, it is necessary to make full use of the sparse power domain characteristic of the non-cellular channel. To reduce the computational complexity dramatically, the factor graph representation method and the belief propagation algorithm have been used in joint signal processing.

The main advantages of cell-free network can be summarized from the following aspects.

- Improvement of system spectrum efficiency: Comparing with traditional cellular network, cell-free network can improve spatial resource efficiency via distributed MIMO antennas. Frequency monochrome multiplexing can be applied into cell-free network, which breaks the principle of frequency multi-color multiplexing in traditional cellular network. In this way, system spectrum efficiency will be promoted largely by cell-free network.
- More flexibility of system coverage: Cell-free Network has higher flexibility in the aspect of system coverage. According to location, number, and requirement of user, beam direction and beam coverage will be adjusted by cell-free network equipped with distributed MIMO antennas. The coverage area of cell-free network will be different from that of traditional cellular network.
- Uniform performance enhancement for users within the coverage area: In the traditional cellular network, users on the edge of the cell usually undergo low transmission performance, which is caused by insufficient network coverage and inter-cell interference. By beamforming flexibly with uniformly distributed RAUs, this problem can be solved in cell-free network with better data rate for users that may otherwise be located at traditional cell edge. In other words, cell-free network provides a higher and more uniform SNR within the coverage area than conventional cellular networks.
- Lower influences in adjacent coverage areas: By adopting flexible beam scheduling, joint signal processing and other methods, the interference between antennas in cell-free network will be reduced through space division multiplexing. Then, the interference in overlapping coverage area can be avoided as much as possible.
- Balance between power efficiency and spectrum efficiency: According to different scenarios, users, and service requirements, cell-free network with distributed antennas or large-scale multiple antennas can make a balance between power efficiency and spectrum efficiency. It makes the entire architecture more flexible.

Currently, the whole concept of cell-free networks has not yet been directly standardized by 3GPP. However, the underlying distributed MIMO technologies has already been studied in the LTE era, and were released as CoMP-JT. Also, the CU/DU split in 5G NR also lays the foundation for realizing the cell-free system. Cell-free network support in the scope of O-RAN was also discussed [66]. After all, as an emerging paradigm, standardization of cell-free networks still faces challenges in terms of technological development and economic viability.

D. Summary of Existing Paradigms

Table IV compares different RAN paradigms. The concepts of C-RAN and HetNet were proposed in 2010s while cell-free network was conceived several years later in 2015. Basically, their objectives are different and independent. Yet, the centralized control paradigm is helpful for enhancing both HetNet and cell-free network. Also, the core rationale behind HetNet and cell-free network is similar, i.e., densifying the cells/APs to increase the average gained capacity of users. The features of C-RAN and HetNet have been partly realized in the current networks, while cell-free network still faces some challenges in real world deployment. In the next section, we

will summarize the trends of cell architecture evolution and show how to integrate the key features and advantages of these existing RANs through the decoupling paradigm and the newly-proposed FD-RAN architecture.

III. Trends of RAN Evolution and FD-RAN

A. Trends of RAN Architecture Evolution

1) Increasing BS Density and Heterogeneity: Increasing the density of BS deployment is an effective means of enhancing network throughput [29]. This leads to smaller cell sizes with users closer to BSs, and allows for spectrum reuse across multiple cells. To further increase density, the coverage areas of cells can overlap. With an increasing number of BSs, heterogeneity may be introduced naturally, meaning that cells can exhibit diversity with regards to size, BS configuration, spectrum, and air interface technology. This allows for BSs with varying characteristics to be utilized in various application scenarios.

2) Enhancing BS Cooperation: The cooperation of different BSs transforms the conventional paradigm of a user being served by only one BS. As BS density increases, multiple BSs can offer services to the user simultaneously, enhancing the user's QoS, especially at the cell edge. There are several methods for BSs to cooperate. The simplest form is multi-connectivity, allowing the user to have multiple connections to different BSs. Usually, a primary connection is used for service, but if its performance drops, another connection can be activated as the primary, ensuring the user's service reliability. More advanced cooperation techniques are belong to coordinated multi-point transmission [67]. BSs can allocate orthogonal frequency resources to the same user or transmit signals to the user on the same frequency using distributed MIMO techniques. Heterogeneity of cells can also be utilized in cooperation, such as using small cells for uplink and macro cells for downlink. To achieve cooperation, two prerequisites must be met: central control of different BSs, and dense and heterogeneous BS deployment.

3) Improving Spectrum Utilization Efficiency and Flexibility: The spectrum is a valuable resource for wireless communication. The use of advanced physical layer transmission technologies, such as beamforming and spatial multiplexing, has significantly improved the spectrum efficiency in bits/Hz. However, in practical cellular networks, the utilization efficiency of the spectrum, i.e., the percentage of the spectrum actually used for data transmission, is often low on both time and space scale. This is due to the lack of flexibility for spectrum utilization in the current RANs. In both TDD and FDD, a portion of the spectrum must be left unused to prevent interference between uplink and downlink. Moreover, without cell cooperation, the unused resources of neighboring cells cannot be utilized to meet the demands of users in the current cell. To increase spectrum utilization flexibility, it is necessary to pool the spectrum resources and adopt centralized scheduling without any predetermined restrictions.

4) Reducing Network Cost: Reducing cost is crucial for the long-term success of cellular networks. Lower costs enable the deployment and operation of more infrastructure and give smaller operators the opportunity to enter the market, leading to improved services for users. Centralizing high-complexity signal processing and placing computing resources at the edge cloud can decrease operational costs. Another effective way to reduce energy costs is to turn off some BSs when total user demand is low [13]. Although increasing BS density and heterogeneity will lead to more capital costs, potential savings can still be achieved during practical deployment and through using reduced capability (RedCap) equipment.

5) Providing Personalized Services: Looking back at the history of cellular networks, we can see that 4G was successful in providing widespread high-speed Internet access to users. Moving forward, 5G is expected to be tailored to specific application scenarios and industries. In line with this trend, we can say that 6G should be able to offer personalized services to users, who usually have diverse requirements. To achieve this goal, BS heterogeneity and cooperation are crucial. The network must also be flexible enough to maintain a high level of quality at any location, so as to meet users' requirements.

6) Becoming AI-Native: Due to the tremendous success of AI models, many research works have employed AI to tackle various challenges in wireless communications. It is believed that AI will be deeply integrated with cellular networks [68], leading to significant transformations, especially with the recent breakthrough of large foundation models. From the standpoint of RAN architecture, the network should first be equipped to gather large amounts of training data and computing resources to facilitate AI adoption. Additionally, the network's operation must be more open and flexible, enabling AI to maximize its potential and enhance performance under various situations.

B. The Decoupling Paradigm and FD-RAN

Considering the advantages of existing network architectures and the trends of evolution mentioned above, we further present a newly-proposed architecture for 6G, namely FD-RAN. The architecture of FD-RAN is depicted in Fig. 8. It is designed with edge cloud assistance, with centralized network control and data processing at the edge cloud. The network serves users through three types of BSs: data BSs, including DL-BSs and UL-BSs, and a C-BS. DL-BSs provide downlink data transmission, while UL-BSs are responsible for uplink data reception. DL-BSs typically require more energy, which is supplied directly from the electric grid, whereas UL-BSs consume less energy, allowing for more flexible deployment with a higher density to save energy for UEs. C-BSs use low-frequency spectrum bands for a larger coverage and mainly handle control message exchange with UEs. Each UE is covered by at least one C-BS, enabling network control and state feedback. It is important to note that C-BS is not used for physical layer channel feedback, which is required for physical layer data transmission in current networks.

At the core network, Cybertwin [69], [70] serves as a new network function for FD-RAN. Cybertwin runs at the edge cloud and serves as a user's entry point to the Internet. Meanwhile, it acts as the anchor of all the user's

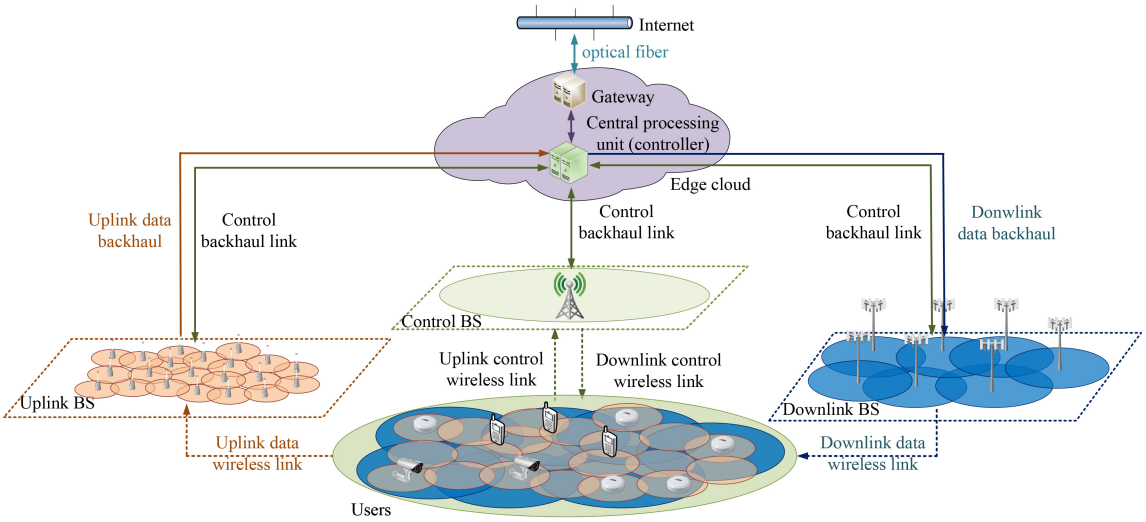

Fig. 8. Architecture of fully-decoupled RAN.

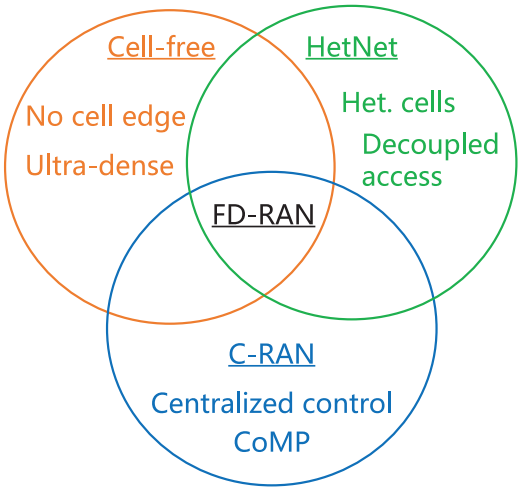

Fig. 9. FD-RAN incorporates key features from C-RAN, HetNet, and cell-free network.

communication links, including other air interfaces such as Wi-Fi and satellite communications. Therefore, it performs multi-connectivity management and mobility management for users. Cybertwin also acts as the interface between users and the cellular network. Through Cybertwin, users can configure their personalized requirements for specific services, enabling the network to satisfy different users' QoE through user-centric resource allocation.

FD-RAN incorporates key features from C-RAN, HetNet, and cell-free network, as shown in Fig. 9. Specifically:

- FD-RAN not only adopts centralized signal processing at the edge cloud as C-RAN to enable cooperative transmission techniques like CoMP, but also enhances the central control in C-RAN by using a dedicated C-BS so as to maintain a dedicated control channel with UEs.
- In FD-RAN, the uplink and downlink BSs are fully separated. Thus, decoupled access in HetNet is supported by default. Also, more heterogeneous uplink and downlink BSs can be incorporated owing to the deployment flexibility introduced by FD-RAN. For example, non-terrestrial BSs can be adopted solely for supplementing downlink. Furthermore, multi-BS cooperative transmission in both uplink and downlink can improve the utilization efficiently of heterogeneous BSs.
- In FD-RAN, uplink BSs are decoupled from downlink BSs, thus it becomes much easier to densely deploy uplink BSs. Then, users' uplink SNR and data rate can also be enhanced uniformly as in cell-free networks, considering that uplink performance is usually the bottleneck in traditional cellular networks. Besides, users can be served in a user-centric fashion in FD-RAN by selecting the most appropriate uplink and downlink BSs, and their personalized service requirements can be satisfied with the help of Cybertwin.

The key features of FD-RAN are further discussed as follows:

1) Separate Uplink/Downlink Transmission: In FD-RAN, the uplink and downlink data transmission are decoupled, and are performed by UL-BSs and DL-BSs, respectively. This allows for a UE to connect with the most suitable BSs for both uplink and downlink transmission. Additionally, a UE can connect with multiple BSs of the same type, which can work together to provide better service. The separation of the uplink and downlink networks means that traditional physical layer transmission methods, which rely on resource-intensive channel feedback, cannot be used. Control-related messages can be transmitted more efficiently through the use of C-BSs and a dedicated control channel, rather than being tightly coupled with data transmission on the same channel.

2) Flexible Spectrum Usage: In FD-RAN, the separation of uplink and downlink networks allows for more flexibility in the use of spectrum. The spectrum is no longer restricted to a fixed use for either uplink or downlink and can be used for any purpose, as long as interference is avoided. Additionally, the ownership of the spectrum can be decoupled from its usage through either spectrum sharing or spectrum trading. In spectrum sharing, operators share their licensed spectrum with

TABLE V
ADVANTAGES AND DISADVANTAGES OF DIFFERENT RAN ARCHITECTURES

6G challenges, new features, and requirements	C-RAN	HetNet	H-CRAN	Cell-free	FD-RAN
BS cooperation and centralized signal processing	Yes	No	Yes	Yes	Yes
UL/DL decoupling	No	Yes	Yes	No	Yes
Improvement of cell edge performance	Partially Yes	No	Partially Yes	Yes	Yes
Flexibility (e.g., spectrum usage, BS deployment and BS on-demand activation)	No	Partially Yes	Partially Yes	No	Yes
Improvement of spectrum efficiency and utilization	Partially Yes	Partially Yes	Partially Yes	Yes	Yes
QoS guarantee/user-centric service	No	No	No	Yes	Yes
Integrating satellite communications	No	Yes	Yes	No	Yes
Signal processing and resource management complexity	High	Medium	High	High	Medium
Fronthaul costs	Medium	Low	Medium	High	Medium
Dense BS/AP deployment costs	Low	Medium	Medium	High	Medium

each other [71], while in spectrum trading, an operator can rent spectrum to another. The contract must specify the valid time and location for using the shared or rented spectrum. This results in more flexible utilization of spectrum resources at both a micro level, between uplink/downlink and different BSs, and a macro level, between operators.

3) User-Centric Network Service Provision: With the always-on control channel between the C-BS and UEs, users' QoE and personalized service requirements can be fed back to the network. This allows the network to become user-centric [20], serving different users in different ways. The flexibility of resource utilization in FD-RAN allows lower-layer network resources, such as BSs, spectrum, and power, to be dynamically scheduled to meet users' personalized requirements. To guarantee QoE and differentiate between users, service-level agreements (SLAs) can be signed with users, and the network should make its best effort to comply with the SLA and compensate users if it fails. As total resources are limited, SLAs with different service qualities should be priced differently [72], with higher-priced SLAs receiving higher priority. Additionally, a user can have different SLAs signed for different applications, based on their value to the user.

4) Reduced Energy Consumption: The most useful technique for reducing energy consumption of BSs is sleeping. However, for fully-functional BSs, sleeping can only be achieved when there is no need to use both uplink and downlink. Instead, in FD-RAN, since the UL-BSs and DL-BSs are physically separated, each BS can be deactivated separately. As a result, there is more possibility to sleep more BSs. Furthermore, the C-BS will be kept always on so as to coordinate the other BSs and UEs, and also to reduce the possibility of coverage holes [73]. For the UE side, it should be clarified that the UE does not use more antennas for maintaining uplink, downlink and control link connections. In FD-RAN, these links use different spectrum. Thus, UE's energy consumption in FD-RAN is similar to FDD system, which is lower than TDD system. Besides, UEs can be associated with the nearby UL-BSs so as to reduce the uplink transmission distance. Uplink receive diversity, signal combining and resource cooperation techniques can also be adopted in FD-RAN. Then, the UE can use less power to transmit data and achieves improved energy efficiency, as showcased in [74].

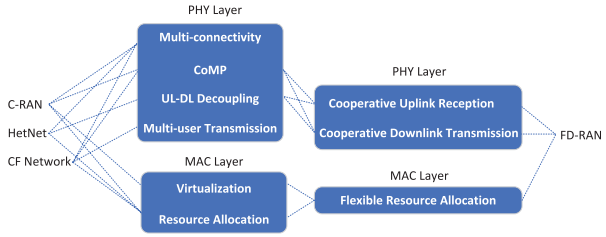

Fig. 10. Relationship between architectures and enabled technologies.

C. Summary of FD-RAN

In summary, the key advantage of FD-RAN is its flexibility in the utilization of network resources, allowing the network to adapt to changes in traffic demand. Specifically, the shortage of spectrum can be alleviated by relaxing its usage and making more spectrum available in areas with high traffic demand. Network costs can be reduced through: (1) on-demand deployment of UL-BSs and DL-BSs, which are cheaper than fully functional BSs; (2) improved spectrum utilization efficiency through spectrum sharing and trading; and (3) increased energy efficiency through techniques such as selective activation of DL-BSs. Personalized service requirements and QoE for users can be guaranteed through user-centric resource allocation, which prioritizes fulfilling SLAs to satisfy users. FD-RAN is also suitable for specific use cases, such as C-V2X [75], [76] and federated learning [77], [78]. In Table V, we also list the advantages and disadvantages of existing RAN architectures as well as FD-RAN. These different RANs are evaluated with respect to various 6G challenges, new features, and requirements. Note that the evaluation results are derived from comparisons among these architectures and are therefore relative.

IV. ARCHITECTURE-ENABLED PHYSICAL LAYER: MULTI-LINK CONNECTION AND TRANSMISSION

In this section, we present survey on multi-link connection and transmission. Each user can connect with more than one BSs, and each connection can incorporate both uplink and downlink or either of them. Fig. 10 illustrates the relationship between architectures and enabled technologies. For example, CoMP can be enabled by C-RAN and CF network, and UL-DL decoupling can be enabled by HetNet. For physical layer

TABLE VI
COMPARISONS OF MULTI-CONNECTIVITY TECHNIQUES

Techniques	3GPP Standardization	Backhaul	Carrier Frequencies	Data Flow Split	Coordination	Performance Gain	Implementation
CoMP	Release 11	Ideal, Fiber-based, High-speed, Low-Latency, Complex and Expensive	Same	PHY Layer	Inter-site and Intra-site	High data rates, imporved cell-edge throughput,	Most Difficult
CA	Release 10	Ideal, Fiber-based, High-speed, Low-Latency, Complex and Expensive	Different	RLC Layer	Inter-site and Intra-site	Capacity enhancement, no reliability enhancement, mobility robustness	Modest
DC	Release 12	Non-ideal, Lower speed, Higher latency, Simpler and Cheaper	Different	PDCP Layer	Inter-site	Improved cell-edge throughput, mobility robustness, reduced signaling, load balancing, energy efficiency	Simple

transmission, through utilizing the advanced signal processing techniques, multiple transmission points can coordinate and transmit to one or more users simultaneously, so as to improve the spectral efficiency and reduce interference.

A. Multi-Connectivity

Multi-connectivity [50], [79] is a general topic as a user can be served by multiple BSs in several different ways. In this subsection, we will cover carrier aggregation (CA) and dual connectivity (DC), which use different carrier frequencies for different links. CoMP will be introduced separately in the next subsection since it works on the same frequency. A comparative analysis of these techniques is presented in Tab. VI.

1) Carrier Aggregation: CA enables users to access BSs with wider bandwidth, resulting in higher data rates, by aggregating multiple independent component carriers from the same or different BSs to form contiguous or non-contiguous frequency bands. The aggregated carriers are classified as primary carriers, which update only during the handover and BS selection process, and secondary carriers, which can be updated anytime. CA supported by 3GPP Release 14 [62] is an intra-site (and inter-band) manner, which has been further extended to a multi-connectivity manner, namely inter-site CA (also referred to as multi-flow CA or multi-stream CA) [54], [80], [81]. In inter-site CA, UEs transmit data simultaneously to MBSs and SBSs in HetNets, resulting in enhanced network performance.

Recent advancements in CA in the context of HetNets are the followings. In [80], the authors employ a queuing analytical model to investigate the cross-layer performance of UEs in HetNets using multi-flow CA. Numerical results demonstrate that the developed model can offload traffic from MBSs to SBSs while guaranteeing the QoS for UEs. An energy-efficient multi-stream CA approach for HetNets is proposed in [54], which is tackled using quasiconvex relaxation. The performance of the scheme is evaluated, and the trade-offs between energy savings and capacity maximization are characterized. A related study that also focuses on optimizing energy efficiency is presented in [56], whereas a novel CoMP scheme enabled by multi-stream CA is proposed. The performance of the proposed scheme is compared with conventional multi-stream CA or CoMP schemes, and it is validated to be superior. Additionally, game-based spectrum allocation [81] and distributed power allocation [82] frameworks are explored to showcase the advantages of multi-flow CA.

2) Dual Connectivity: Challenges such as mobility robustness, per-user throughput enhancements, and increased signaling load due to frequent handover have prompted the development of DC. DC was first specified in 3GPP Release 12 [60] to allow users to obtain resources from both the Master BS (commonly MBS in HetNets), which serves as the main resource provider and mobility anchor, and the Secondary BS (commonly SBS in HetNets), which serves as a supplementary resource provider, simultaneously using the same or different carrier frequencies. DC is analogous to inter-site carrier aggregation techniques [83]. However, DC aims to leverage the advantages of inter-site carrier aggregation by non-ideal backhauls instead of fiber-based connections between macro and small BSs.

In a study by [60], the effectiveness of dual connectivity at the same carrier frequencies was investigated, demonstrating that it can effectively increase the throughput of micro users by reducing interference from macro BS. However, it may not be valid for macro users, as the interference from micro BS is relatively slight. In [84], the authors summarized the architecture and functionalities of DC from the perspectives of the user and data plane. They exemplified the dual connectivity DL/UL user plane protocol stacks and mobility management procedures to explain how DC operates. System-level simulations also demonstrated the superiority of DC over single connectivity in terms of throughput and mobility, with acceptable performance degradation compared to inter-site carrier aggregation. In [85], high-mobility vehicular networks are considered, and the authors leverage multi-connectivity to guarantee the uRLLC service. A deep reinforcement learning method is developed to realize power allocation for multi-connectivity uRLLC.

The evolution to 5G is a gradual migration from LTE, with consideration given to capital costs and backward compatibility. DC introduced in LTE can be exploited to achieve a smooth transition, which not only addresses coverage holes in the early stages of 5G deployment but also provides a viable and cost-effective solution for network operators. 3GPP has proposed different architecture options for 5G deployment, including the transition from legacy LTE to full-fledged 5G. However, only some of these options are suitable for implementation in the

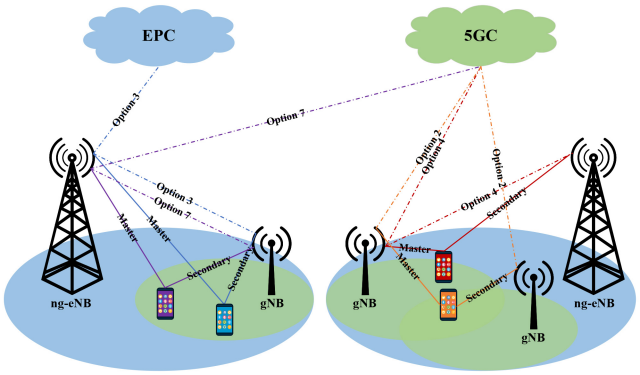

Fig. 11. Different MR-DC configurations in HetNets.

transition stage, namely Option 3, 4, and 7, while Option 2 represents the ultimate goal of 5G deployment. These different options lead to different multi-radio dual connectivity (MR-DC) configurations, as defined in [61] and illustrated in Fig. 11:

- EN-DC: The user is connected to both LTE and NR, where the ng-eNB is the master BS and gNB is the secondary BS.
- NGEN-DC: The user is connected to both LTE and NR, where ng-eNB is the master BS and gNB is the secondary BS.
- NE-DC: The user is connected to both LTE and NR, where gNB is the master BS and ng-eNB is the secondary BS.
- NR-DC: The user is only connected to NR, where one gNB is the master BS and another gNB is the secondary BS, or one gNB is both the master and secondary BS.

The differences also lie in their core network utilization. In the EN-DC configuration, the evolved packet core (EPC) is utilized, while 5G core (5GC) is used in the other configurations. It should be noted that the first three configurations are unique in the sense that DC is adopted to connect two different generations of RAN.

Based on the aforementioned definitions and Fig. 11, it can be observed that EN-DC aligns with Option 3, NGEN-DC corresponds to Option 7, NE-DC can be associated with Option 4, and NR-DC can be attributed to Option 2. An overview of these architecture options is provided in [86], [87]. For EN-DC, specifically Option 3, 3a, and 3x, [88] provides a comprehensive introduction to the architecture, data plane, and control plane functions, as well as a comparison of these options in terms of data rate, user plane and control plane latency, mobility and reliability, and beam failure recovery. The dual connectivity operation in NE-DC is analogous to EN-DC, and thus discussion relevant to NE-DC can consult EN-DC. However, NGEN-DC requires more improvements on legacy LTE BSs compared to NE-DC [89]. Furthermore, NR-DC is more in line with traditional LTE DC but exhibits more characteristics of NR, such as high throughput, high isotropic pathloss, poor diffraction, and more [90].

B. CoMP

CoMP in 6G is a promising solution that improves cell edge completion while minimizing coordination complexity [67].

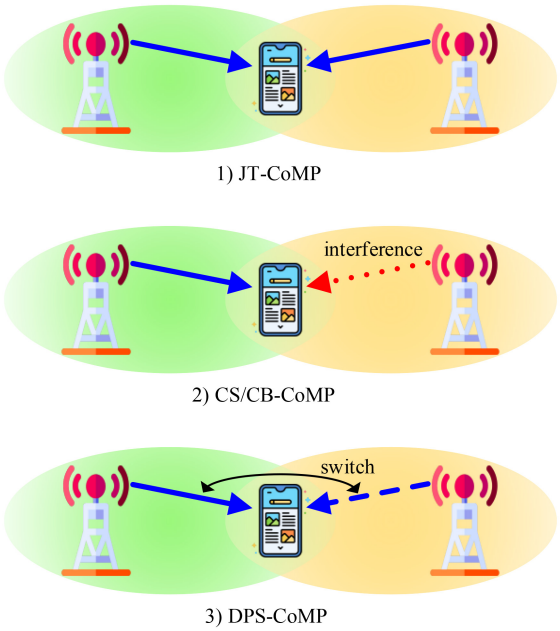

Fig. 12. Three types of CoMP.

Combined with large bandwidth, Massive MIMO and artificial intelligence, the potential of CoMP is further unlocked by fully optimizing and coordinating concurrent transmissions from multiple RRHs, the efficiency, throughput, and coverage of the multi-cell system can be greatly improved. However, coordinated cells require precise synchronization and bring additional scheduling, signal processing, pilot overhead, and complex beamforming design [91]. Fortunately, virtualizing BBUs into a BBU pool provides the necessary conditions for achieving CoMP, and enables new technologies to be combined with CoMP [92].

It was firstly identified as one of the key functions for physical layer enhancements in the 3GPP research project on LTE-Advanced [93], [94]. There are various types of CoMP coordination schemes, and 3GPP has identified three main downlink coordination categories for LTE-A [95], [96], [97]. With Fig. 12, these three types of downlink CoMP are illustrated as follows.

1) Joint Transmission (JT): In JT-CoMP, a large distributed single antenna array is formed by the antennas of a coordinated RRH cluster [98]. This allows the same data to be simultaneously transmitted to the UE by all coordinated RRHs, resulting in significantly improved communication performance. Multiple RRHs can serve a single user in a coherent manner, where interference signals are converted into useful signals. Coherent transmission is achieved through synchronized transmission and joint precoding, which ensures that the signals from different coordinated transmission points are co-phased at the transmitter via precoding. While this type of coordination scheme offers the best results, it requires a high backhaul bandwidth with low latency due to the need for extremely precise information synchronization, including CSI and scheduling, among multiple coordinated RRHs.

2) Coordinated Scheduling/Coordinated Beamforming(CS/CB): In CS/CB-CoMP, a single RRH transmits the signal to the UE while other RRHs in the coordinated cluster use scheduling and beamforming functionality to dynamically mitigate inter-cell interference [99]. The CSI of all RRHs is first shared in the BBU pool for coordinated scheduling/beamforming operations, and then the transmit data is sent to the serving RRH. Beamforming between different RRHs can reduce interference, and scheduling can help mitigate strong interference conditions. Due to the reduced data exchange, CS/CB requires lower backhaul bandwidth compared to JT-CoMP.

3) Dynamic Point Selection (DPS): In DPS-CoMP mode, user data is transmitted from a single serving RRH, while the serving RRH can be dynamically changed to another RRH in different subframes depending on channel conditions and resource availability among a set of available RRHs within the cluster. In each transmission interval, all available RRHs in the CoMP cluster check their channel quality to the UE and dynamically select the RRH with the maximum channel gain for data transmission. Compared to JT-CoMP and CS/CB-CoMP, DPS-CoMP offers a good trade-off between system performance, transmission algorithm complexity, and backhaul overhead.

CoMP techniques are also utilized in the context of HetNets, that is, the Scenario 3 of CoMP scenarios proposed by 3GPP [96], which utilizes low-power remote radio heads for expanded coverage. CoMP is more significant and effective in HetNets than in homogeneous networks due to the considerably higher interference. While the principal ideas of CoMP and eICIC are similar, CoMP outperforms eICIC in terms of performance and complexity. The differences are listed as follows [67]: Furthermore, some joint CoMP and eICIC schemes in HetNets are investigated in [100], [101], [102]. Considering the increasing energy consumption of densely deployed HetNets, [103] and [104] propose energy-efficient CoMP schemes assisted by simultaneous wireless information and power transfer (SWIPT) for green HetNets. The proposed NOMA-aided CoMP framework for HetNets, as presented in [105], leverages JT-CoMP and CS-CoMP techniques for less distinctive channels with multiple BSs, whereas non-CoMP is employed for the dominating channel with a single BS. Numerical results attest to the framework's superior spectral and energy efficiency in comparison to the conventional CoMP-orthogonal multiple access scheme. A similar scheme [106] is investigated in virtualized HetNet aiming at reducing the complexity of users' successful interference cancellation.

C. UL-DL Decoupling

Traditional networks have primarily focused on DL performance because of the dominance of DL services [107]. However, due to the emergence of new technologies, such as cloud techniques, machine-type communications (MTC) networks, dense HetNets, and increasing subscribers and devices, UL traffic is anticipated to escalate rapidly [108]. As a result, maintaining UL performance has become a significant area of research interest.

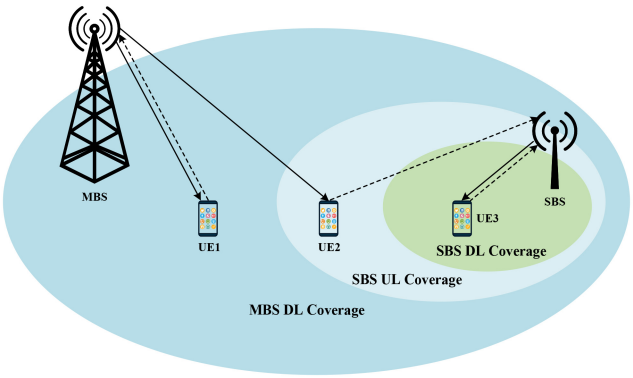

Fig. 13. Decoupled UL-DL association in HetNet.

Despite the growing importance of UL, improving its performance remains a challenge due to the inherent imbalance between UL and DL transmissions. BSs can utilize more antennas and higher transmit power to compensate for path-loss, while the transmit power and antennas of UEs are quite limited. Moreover, the gap between UL and DL coverage is further exacerbated by higher frequency bands, resulting in more severe path-loss, and massive MIMO techniques that provide more multiplexing and diversity gains on the BS side, but not on the UE side [109].

In addition to inherent limitations in user equipment, another significant challenge to UL performance stems from HetNets. In homogeneous networks, users connect to BSs based on the maximum DL reference signal received power (RSRP) in both UL and DL, known as coupled UL-DL access. However, this approach is not appropriate for HetNets. As shown in Table II, MBSs have a higher transmit power than SBSs, resulting in a significant number of users accessing MBSs. In contrast, in UL, users prefer BSs that are geographically closer, i.e., SBSs, since the transmit power of UEs is uniform, and the received power is largely determined by the distance between the user and the BS. This contradicts the access strategy in DL and leads to poor UL performance, particularly for cell edge users. Moreover, denser deployment of BSs exacerbates this poor performance [110]. Although techniques such as cell range extension (CRE) can alleviate this problem to some extent, high interference at the cell edge remains an issue [111].

In response to the aforementioned challenges, the UL-DL decoupling framework has been proposed in literature [112], [113], [114] with the emergence of HetNets. This framework allows for more superior UL and DL associations that are determined by path-loss and received power, respectively. In Fig. 13, we illustrate the decoupled UL-DL association in a HetNet. Specifically, UE2 is located within the downlink coverage area of the MBS but beyond the SBS. As a result, UE2 should connect to the MBS in both downlink and uplink as per the coupled association manner. However, as UE2 is also within the uplink coverage area of the SBS, it can achieve better performance by associating with the MBS in downlink and the SBS in uplink, respectively, i.e., in a decoupled association manner.

The advantages of UL-DL decoupling arise from the optimal UL and optimal DL associations, which ensure that each transmission is directed towards the most suitable BS. Decoupled access to serving BSs leads to a shorter distance, resulting in lower path-loss, transmit power, interference, and consequently higher signal-to-noise-and-interference ratio (SINR). Additionally, UL-DL decoupling allows for more traffic offloading from MBSs to SBSs, leading to a reduction in the burden on MBSs and an improvement in the utilization of SBSs.

The concept of UL-DL decoupling has been standardized in 3GPP Release 12 [60] utilizing DC, and preliminary performance evaluations demonstrate significant performance gains, particularly at the cell edge. In [57], the authors present five arguments in support of UL-DL decoupling, highlighting its potential performance benefits and practical deployment. The paper also outlines the enabling architectural frameworks and future research directions in this domain. Since its inception, UL-DL decoupling in HetNets has been extensively studied in the literature to demonstrate its superiority. Theoretical performance analyses have investigated both spectral efficiency [115], [116], [117] and energy efficiency [115], [118], [119] to compare decoupled UL-DL with coupled access. Studies have shown that decoupled access can provide significant improvements in load balance, fairness, and overall system capacity [115], [116]. The spectral efficiency of HetNets with mmWave BSs and microwave BSs has also been evaluated [117]. In addition, other studies have investigated the benefits of UL-DL decoupling in scenarios such as non-uniform user distribution [120], full-duplex operation [121], ultradense networks [110], multi-tier HetNets [122], multi-antenna BSs [123], and multiuser MIMO [116]. Some studies have also investigated the latency gain in decoupled UL-DL [124], [125] and proposed signaling mechanisms for decoupled UL-DL networks [126]. Furthermore, an overview of association policies in UL-DL decoupling networks is provided in [127], and a recent survey provides a holistic review of operating band UL-DL decoupling at the same BSs [30].

D. Multi-User Transmission

Low-complexity detection is the key to multi-user transmission technology. In centralized MIMO system, the channel correlation matrix is mainly used to reduce complexity. While the correlation of distributed MIMO channel is low, other methods are needed. The precoding technology can effectively eliminate interference among multiple users and significantly improve the system's capacity. At the same time, the appropriate precoding method can simplify the receiver algorithm, which is an important way to improve the system's performance.

1) Precoding: In [128], to suppress multi-user interference completely, the block diagonal (BD) precoding was proposed. However, BD precoding method can suppress multi-user interference at the cost of noise enhancement. In order to overcome this shortcoming, the generalized zero-inverse (GZI) matrix method was proposed in [129], whose sum rate increased linearly with the number of users and transmitted antennas. Its complexity was lower than BD precoding. Further, the authors in [130] proposed a low-complexity linear multi-user multi-input multi-output Gyre precoding based on gradient descent algorithm. This method could find the rotation angle with low complexity and achieve a throughput gain of 13% higher than that of forced zero precoding method. In [131], it aimed to solve the performance degradation problem in multi-carrier, multi-user large-scale MIMO systems, when Doppler expansion was enormous. Using an OTFS-based multi-user pre-encoder at the BS, the proposed LCD detector were used at each user terminal (UT). All the message symbols of UT were jointly demodulated. The analysis results showed that as the number of BS antennas increases, the performance of the proposed LCD detector will achieve to that of the optimal detector. In high mobility scenarios, the proposed OTFS multi-user large-scale MIMO pre-encoder achieved a better SE than that of OFDM-based multi-user large-scale MIMO pre-encoder.

In [132], the authors considered the problem of communication over repeater-assisted MIMO channels and designed three types of rational forced pre-encoders. By formalizing the precoder design problem, computation complexity and system performance can be compromised. In [133], the Space-Time coding approach was adopted to align the interference in the multi-user system. The left null space of channel matrix was used to detect the information symbols at different receivers. This new transmission strategy can save the transmission time-slot resources significantly.

2) Detection: In super-large scale MIMO (XL-MIMO) systems, the computational cost of multi-antenna processing will increase, and the average energy of a given user's signal will varies throughout the array. In [134], it proposed a distributed receiver based on variational message passing (VMP) to solve above problems. The receiver had a flexible complexity-performance tradeoff architecture. Performance was close to that of a genie assisted receiver by utilizing spatial non-stationary properties. The complexity was lower than that of traditional central linear processing methods, like ZF precoding method.

For the unsolicited random access problem of multiple receiving antennas, the user was restricted by using the same codebook and cannot be distinguished. In [135], it proposed to divide user messages into two parts. The pilot portion was selected from the joint codebook, and standard block codes encode the message bits. The multi-measure vector approximate message passing (MMV-AMP) algorithm was used to estimate the active user channel from the pilot portion. Then, the estimated channel was used to perform a coherent maximum ratio merge (MRC) to decode the information bits. The results showed that the proposed scheme was superior to the existing schemes in terms of energy efficiency and much lower complexity when the coherent block length was significantly larger than the number of active users.

In [136], an iterative solution for transmitting/receiving beamforming with total digital minimum mean square error (MMSE) was derived in a millimeter-wave full-duplex MIMO system. On this basis, the RF and baseband filters close to the

complete digital beamforming were designed by the orthogonal matching (OMP) algorithm, and the digital beamforming matrix was derived by the BD method to reduce intra-user and inter-user interference. Authors in [137] proposed two kinds of the semi-blind receiver based on tensor model. The most prominent feature of this receiver was that it can estimate channel state information without pilot sequence and had higher spectral efficiency. The simulation results showed that the proposed receiver had a higher bit error rate (BER) and normalized mean square error (NMSE) performance than the existing receiver under the two modes of TDD and FDD. In [138], the authors put forward different views on the previous MMSE receiver. The actual MMSE receiver not only depended on the CSI error statistics, but also on the autocorrelation coefficient of the time-varying channel. By deducing the quasi-sealing formula of SINR, the MMSE receiving function can reach a higher SINR. In [139], the closed-form expression of mean vector symbol error rate (VER) for packet detection of maximum likelihood (GD-ML) receiver was derived. It was proved that GD-ML was superior to the forced zero (ZF) and MMSE receivers in terms of VER. Moreover, GD-ML receivers had almost the same complexity as ZF and MMSE receivers regarding floating point operations.

Unlike the above ideas, [140], [141] applied the artificial intelligence algorithm to the receiver to achieve higher performance. Reference [141] proposed a machine learning (ML) enhanced MU-MIMO receiver based on the standard linear least mean square error (LMMSE) architecture. It retained the interpretability and extensibility of the LMMSE receiver while improving accuracy. Reference [140] proposed the limit learning receiver based on machine learning. By appropriately setting the number of hidden neurons, ELM achieved higher spectral efficiency and lower bit error rate with fewer floating point operations. In [142], it focused on analyzing the interrupt performance of four wide linear (WL) receivers and proved that WL receivers had higher diversity gain and user reuse capability.

In order to further reduce the complexity of channel detection, [143] adopted confidence propagation detection based on factor graph in MU-MIMO system encoded by low-density parity check (LDPC), which was called FG-GAI BP detection. The proposed iterative joint detection and decoding (JDD) strategy can make the detection scheme converge quickly and achieve a lower bit error rate. In a distributed MIMO system, the downlink and the receiver model after uplink interference suppression were typical linear models under colored noise. In [144], the iterative soft-decision interference elimination (ISDIC) method was used to suppress the inter-user interference. In frequency-selective channels, ISDIC had low complexity and excellent performance.

E. Lessons Learned

From the review of multi-link connection and transmission technologies, we can learn the following lessons so as to provide guidance to the development and design of a different RAN for 6G:

- It is suggested that 6G RAN architecture should have dedicated control infrastructure and resources, so as to enable reliable, low-latency, and large-scale control of multiple BSs and UEs. With the independent and enhanced control functionality, not only signal processing can become more resource-efficient and cost-efficient due to centralization, but also more transmission opportunities can be explored. For example, although multi-user transmission techniques such as MU-MIMO can be realized within a single cell, it is less efficient due to the difficulty in selecting proper users from the same cell. However, with the global information and control of UEs through the dedicated control channel, there will be more chances to carry out multi-user transmission for a certain BS. Dedicated control also paves the way for realizing flexible cooperation among BSs, e.g., via CoMP.
- It is further suggested that UEs' uplink and downlink should not be bound with certain BSs or cells for data transmission. It is obvious that UEs will be in a multi-link environment consisted of multiple heterogeneous BSs. Thus, these links need to be flexibly selected depending on the UEs' situations and the properties of links, so as to efficiently utilize them. The decoupling of uplink and downlink for the same UE can further improve performance, especially for uplink transmission. Moreover, the heterogeneous BSs and resources can be tailored for either downlink or uplink.

V. Architecture-Enabled MAC Layer: Resource Virtualization and Allocation

In this section, we present surveys on resource virtualization and allocation. First, the aggregated various kinds of network resources should be pooled. Virtualization technologies will play an important role. Then, the pooled resources need to be efficiently utilized, which can be realized through appropriate resource allocation. Since resource allocation is a large topic, and strongly depends on the scenario and system architecture, we will survey the literature on resource allocation in different RANs, respectively.

A. Virtualization

The remarkable evolution of the RAN in the past three decades has led to a diverse and complex infrastructure, making it increasingly challenging to develop new network technologies. Traditional BBU processing boards are designed for specific standards and support only a fixed number of carriers, which may cause issues during the rapid development of wireless technologies and standards [39]. The addition of virtualization technology can greatly reduce the CapEx and OpEx associated with these challenges [145], [146]. Virtualization is considered a promising technology for RAN evolution. Recently, virtualization has gained significant attention in the communications and networking fields. For instance, in order to merge user-centric and service-centric networking, holistic network virtualization is proposed from both the service demand and provision perspectives, with its main technologies including digital twin, artificial intelligence, and

network slicing [68]. In this subsection, we first discuss the virtualization of various kinds of physical network resources. Then, we introduce the virtualization of network functions at a higher level.

1) Network Resource Virtualization: Virtualization in C-RAN offers significant benefits for the efficient deployment and operation of wireless networks. It abstracts physical entities, spectrum resources, and network functions into logical units, enabling allocation and utilization that maximizes performance while guaranteeing user requirements [147]. This approach offers several advantages, including improved hardware utilization, decoupling from the network infrastructure, easier migration to newer technologies, and flexible network management. Consequently, it leads to more efficient utilization of network resources and reduces the cost of deployment and operation of wireless networks [148].

The virtualization in C-RAN has a quite broad scope ranging from BBU pool, RRH and radio frequency resources. In general, according to the architecture of C-RAN, the virtualization of C-RAN can be divided into three main components: computing hardware, network facilities and radio frequency resources.

Computing hardware virtualization is a powerful technique that abstracts isolated computing units and aggregates them into a pool of computing resources. It provides users with a unified and abstracted computing platform, while hiding the physical characteristics of computing units. This technique can be applied to communication systems to create BBU pools, which can address the problems associated with fixed designs. A range of hardware components, including network interface cards, memory, CPUs, and storage units, are virtualized using a software platform implemented by the operating system [149]. C-RAN BBU virtualization provides numerous advantages, such as reduced capital investment, lower costs, reduced power consumption, and increased flexibility and reliability in utilizing the computing resource.

The network facility in C-RAN, which includes the fronthaul and RRHs between the BBU pool and mobile users, can also be virtualized to simplify management and optimize the wireless network. Virtualization allows the RAN controller to focus on the capabilities and loads of the facilities in terms of communication and radio frequencies, rather than their physical differences and diversity. By efficiently utilizing the limited fronthaul bandwidth to transmit IQ data based on real-time load information, the controller can maximize the utility of the fronthaul [150]. Virtualizing independent RRHs into a collaborative system enables a distributed antenna to support advanced technologies such as CoMP. The benefits of network facility virtualization include improved performance, reduced cost, and increased flexibility.

Radio spectrum is a critical resource for wireless communications, comprising licensed and dedicated free spectrum [151], [152]. Virtualizing spectrum resources enables the C-RAN controller to gain a comprehensive view of available communication frequencies. This allows for more flexible and efficient use of fragmented spectrum, dynamically allocating frequencies to users based on their real-time demands. Additionally, it facilitates cooperation among telecommunication operators, allowing multiple contracted operators to use all or part of licensed spectrum under their cooperation agreements for spectrum sharing. This results in more efficient and effective spectrum usage for wireless communication.

2) Network Function Virtualization: The advancements of C-RAN are founded on its capabilities, such as programmability, softwarization, virtualization, redesigning radio interfaces, and resource coordination. These improvements are made possible by the enabling technologies of software-defined networking (SDN) and network function virtualization (NFV) [153], [154]. SDN achieves decoupling of control and data planes, while NFV abstracts functionalities from the underlying hardware.

The separation of control and data planes in networking is a key trend that can be leveraged to enable wireless virtualization in C-RAN [31]. By breaking vertical integration, SDN simplifies policy enforcement and network evolution through the use of a logically centralized controller and basic forwarding devices, such as switches and routers [155]. In the context of C-RAN, switches can be equated to BBUs in a pool, while controllers serve as pool coordinators that manage these BBUs.

NFV is a promising technique that aims to consolidate various network equipment devices into software running on high-volume servers, storage units, and switches in the cloud. By leveraging standard computing virtualization technology, NFV reduces costs, preserves the power and physical footprint of the equipment, and makes it easier to adapt to changing network needs. In the context of C-RAN, NFV virtualizes the computational resources while clouding the physical radio resources. This allows BBUs to be implemented as virtualized network functions, providing greater flexibility and agility in the management of wireless networks [156], [157].

To summarize, one of the key advantages of virtualization is that it allows network services to be decoupled from their providing infrastructures. This means that multiple differentiated services can share the same infrastructure, maximizing utilization, and reducing the number of required infrastructures. This approach is expected to save operators up to 40% of the $60 billion spent on OpEx and CapEx over a five-year period [158]. Additionally, virtualization facilitates migration to newer technologies or protocols while still supporting legacy products. By isolating a portion of the network, multiple experiments can be launched and operated concurrently, allowing experimental functionality to be evaluated and deployed without disrupting normal services, even in a real infrastructure [159]. Finally, wireless network virtualization can provide a powerful and convergent network management mechanism for the emerging complicated wireless networks. The BBU pool of C-RAN can dynamically and efficiently control multiple RRHs over a large area, enabling extensive and sophisticated collaboration that results in better network performance.

B. Resource Allocation in C-RAN

Effective resource allocation in C-RAN can address various challenges related to latency, capacity, and connectivity.

Furthermore, the convergence of C-RAN and deep learning is anticipated to bring new possibilities for academic research and industrial applications. Deep learning is a suitable solution for enhancing data processing capacity, cloud resource management, and cellular communication traffic prediction [32]. In the following, we will cover task scheduling, RRH selection, spectrum management, power allocation, and joint optimization [160].

1) Task Scheduling: Task scheduling involves selecting a group of communication tasks to be executed at a specific time period based on channel conditions and available computational resources [161]. Given the limited resources and interference constraints, intelligent scheduling of tasks is essential for improving network throughput and minimizing interference. In [162], existing scheduling approaches suitable for C-RAN are studied, and their potential limitations are identified. To address these issues, a hierarchical scheduling framework is proposed to mitigate the negative impact of fronthaul delays on the throughput of non-cell edge users and enable efficient retransmission of erroneous data, while still benefiting cell edge users from interference mitigation techniques that require centralized control. Additionally, to improve network utilization and reduce deployment costs, a flexible scheduling algorithm for C-RAN architectures is proposed to replace demand forecasts that are uncertain and do not reflect actual fluctuations [163]. Besides, a load balancer called Dynamic Greedy Spike (DGS) is designed and implemented for a C-RAN architecture using user-level virtualization, allowing network operators to assign new users to any host in the cloud, regardless of their source base station [164]. By modeling the issue as weighted improper graph coloring, DGS assigns users to different hosts to improve isolation and decrease user interference.

2) RRH Selection: The selection of RRHs is a critical task that directly impacts network spectral and energy efficiency. Moreover, RRHs can collaborate and perform centralized beamforming tasks, enhancing wireless channel throughput. To mitigate the network performance decline caused by the unnecessary handovers and blocked users, an RRH-Sector pair selection for new connections and network load-balancing framework is proposed to optimize the network performance and operator reward without affecting the users QoS in C-RAN [165]. The network power consumption for a user-centric is minimized by jointly optimizes the precoding matrices and the set of active RRHs, where both users' rate requirements and per-RRH power constraints are considered [166]. This problem is tackled in two stages: first, a low-complexity user selection method is proposed to obtain the largest subset of feasible users, followed by a low-complexity algorithm to solve the network optimization problem with the selected users. In order to avoid the long processing delays and high computational burden associated with optimizing different resources in different network dynamics, a supervised deep learning technique is proposed to ensure the QoS requirements of users [167]. It solve the joint resource allocation and RRH-association problem in multi-tier C-RAN, where an efficient RRH-association, sub-channel assignment and power allocation technique are applied to generate training data.

3) Spectrum Management: Spectrum management is the key to reaping the true benefits of C-RAN technology. An efficient multi-class classification radio resources management scheme based on the cooperative evolution of support vector machine (SVM) is proposed to assign sub-channel of variable bandwidth to RRH users and D2D pairs, such that sub-channels can be reused without compromising the QoS requirement [168]. A resource allocation strategy based on the effective bandwidth and 5G physical layer enablers together with CoMP is presented to meet the reliability and latency requirements of ultra-reliable low-latency communications (uRLLC) traffic under the restrictions of fronthaul capacity and RRH resource availability [169]. The packet delivery method, time-frequency resource allocation, and queuing strategy for the CoMP enabled uRLLC in C-RAN are investigated, where the interplay between the error components is used to minimize bandwidth and the end-to-end error and delay components are identified. An network slicing algorithm is studied for C-RAN than incorporates two services: enhanced mobile broadband (eMBB) with multicasting for improved throughput and uRLLC with finite blocklength capacity to accurately capture delay [170]. Its goal is to maximize revenue by admitting slice requests while meeting limited resource constraints, and the problem is formulated as a mixed-integer nonlinear programming problem and solved using efficient methods such as semidefinite relaxation and successive convex approximation.

4) Power Allocation: Effective power allocation is a critical aspect of wireless networks, but it is particularly challenging in C-RANs due to the close proximity of RRHs and resulting interference issues. Furthermore, proper power allocation is crucial for achieving high energy and spectral efficiency. To improve the uRLLC reliability of vehicular network, a multi-connectivity uRLLC downlink transmission scheme is studyed and a multi-agent cooperative deep reinforcement learning algorithm, called transformer associated proximal policy optimization (TAPPO), is proposed to achieve real-time robust power allocation for multi-connectivity uRLLC with imperfect CSI [85]. To support different use cases with diverse QoS, the efficient resource allocation problem in C-RAN is formulated as a mixed integer nonlinear program, and an algorithm based on penalized successive convex approximation of polynomial time complexity is designed to determine a sub-optimal solution [171]. A resource allocation method for C-RAN is developed for the optimization issue that is intended to maximize the network's energy efficiency, subject to practical restrictions such as QoS requirements, RRH transmit power limits, and fronthaul capacity limits [172]. This non-convex, mixed-integer network energy efficiency maximization problem is transformed into a weighted sum-rate (WSR) maximization problem through successive convex approximation methods and solved through a provably-convergent iterative method.

5) Joint Optimization: Optimizing multiple parameters simultaneously through joint optimization is a powerful and practical approach in C-RAN. However, it is also one of the most challenging approaches due to the large number of parameters that need to be optimized. The dynamic user

TABLE VII
STATE-OF-ART LITERATURE REVIEW ON USER ASSOCIATION, RESOURCE BLOCK ALLOCATION, AND POWER ALLOCATION IN HETNET

Catergories	Ref	Optimized Resources	Performance Metrics	Algorithms	Operations
Optimization-Based	[189], [179], [180], [181], [182]	User Association [180], [181] Carrier Allocation [178] Subchannel Allocation [179] Bandwidth Allocation [180] Power Allocation [178], [179], [180], [181] Time Allocation [182]	Energy Efficiency[178], [179], [180], [181] Overall Delay[182]	[178] Convex Alteration Estimation, Fractional Programming, Lagrange Dual Method [179] Relaxation Approach, Worst-case Transformation [180] Relaxation Approach, ADMM [181] Fractional Programming, Two-phase ϵ-optimal outer-approximation [182] Layered Algorithm, Myopic-gradient	Centralized [178], [179], [181], Distributed [180], [182]
Optimization-Based Combining Others	[183], [184], [185], [186], [187], [56]	User Association [185], [186] Computing Resource Assignment [183] Channel Reusing [183] Edge Resource Allocation [183] Subchannel Allocation [184], [185], [56] Bandwidth Assignment [186] Precoding Vector [187] BS Clustering [56] Power Allocation [184], [185], [186], [187], [56]	Energy Consumption [183] Energy Efficiency [184], [56] Network Cost [185] User Rate and Cross-Tier Interference [186] Sum Rate of All Users [187]	[183] Lagrangian Multiplier Method, Matching Theory, Greedy-based Algorithm [184] Lagrangian Dual Method, Modified Gale-Shapley Matching [185] Alternative Search Method, Hungarian Algorithm, SCA [186] Swapping based Algorithm, Closed-Form Derivation, SCA [187] Fractional Programming, Quadratic Transform, Genetic Algorithm [56] Sequential Minimization Technique, Zoutendijk Feasible Direction Method, Genetic Algorithm	Centralized [183], [184], [185], [186], [187], [56] Distributed [183]
Game Theory	[188], [189], [190], [47], [191]	User Association [188], [191] Spectrum Allocation [188] Computing Resource Management [189], [47] Beam Allocation [190]	Efficiency and Load Balancing [188] Maximum Integral Utilities with Stable User Distribution [189] Energy Efficient Beam [190] Utilities for Sellers and Buyers [47] Efficient Association [191]	[188] Non-Cooperative Game [189] Evolutionary Game, Stackelberg Differential Game [190] Cooperative Game [47] Two-Stage Auction Game [191] One-to-Many Matching Game	Centralized [189], [190], [47], [191] Distributed [188]
DRL	[192], [193], [194], [195]	User Association [193], [194], [195] TDD Configuration [192] Spectrum Allocation [193] Power Allocation [193], [194], [195]	Data Transmission Rate [192] Network Rate [193], [194] Proportional Fairness [195]	[192] Deep Belief Neural Network, DQN [193] Distributed Coordinated Multi-Agent DDQN [194] Single-Agent DQN, Multi-Agent DDPG [195] RNN, Closed-Form Policy Gradient	Centralized [192], [194], [195] Distributed [193], [194]

scheduling and power allocation problem is formulated as a stochastic optimization problem with the objective to minimize the total power consumption of the whole network, and it is solved by transforming it into a series of static optimization problems [173]. To accomplish the minimum total system transmit power with low complexity algorithm subject to sparse code multiple access (SCMA) restriction, total available power in RRH, fronthaul limitation, and QoS prerequisites for each user, a downlink joint codebook assignment, user association, and power allocation for a SCMA-based systems in C-RAN are considered [174]. A joint energy minimization and resource allocation problem in C-RAN with mobile cloud computing is solved by an iterative algorithm, where the constraints includes task executing time, transmitting power, computation capacity and fronthaul data rates [175]. A joint resources allocation scheme is studied for network slicing in C-RAN, which combines network slicing with C-RAN and solves the joint spectrum and computing resource allocation problem by modeling it into a MINLP problem and then decomposing it into two subproblems [176]. A machine learning scheme called Twin-GAN-based DRL (TGDRL) is proposed to jointly allocate both wireless bandwidth and computational resources by extending existing GAN-based DRL results [177]. This scheme employs a multi-objective optimization algorithm to improve spectrum efficiency and reduce computational resource consumption, which simultaneously addresses both bandwidth allocation and computational resource allocation.

C. Resource Allocation in HetNet

In HetNets, the increased network density and the presence of diverse multi-RATs allow users to access a more extensive selection of BSs, providing more degrees of freedom in resource scheduling. However, this also presents challenges for resource allocation, as more ubiquitous BSs are constrained by limited resources, resulting in more complicated management than in previous networks. To this end, we present recent advances of resource allocation in HetNets, focusing on two aspects: user association, resource and power allocation; and interference management.

1) User Association, Resource Block and Power Allocation: Apart from interference management, resource allocation can be classified into the following parts: user association, resource block allocation, and power allocation. User association involves establishing connections between users and BSs, which can be many-to-one or even many-to-many. Resource block allocation is defined as the scheduling of physical resource blocks based on established connections, while power allocation involves the allocation and control of power. These three parts are interdependent and indivisible, and are often jointly optimized. We are concerned with approaches that tackle these parts to form an efficient resource allocation and do not distinguish between them particularly. We provide a comprehensive summary of the state-of-the-art literature in Tab. VII, categorizing the studies into four distinct groups based on their methods: optimization-based, optimization-based combining others, game theory, and DRL. Additionally, key characteristics such as optimized resources, performance metrics, algorithms, and operations are detailed in the table.

In order to develop green HetNets for the future, authors in [178] aim to minimize overall energy consumption by optimizing carrier allocation and power utilization, which in turn determine whether small cells should be put to sleep. In the energy harvesting (EH)-enabled HetNet that incorporates a MBS and multiple SBSs, a fractional problem is formulated to

optimize the energy efficiency. To address the non-convexity of this problem, the authors estimate a lower bound of user throughput and apply parametric programming to transform it to a convex form. They then employ the Lagrange dual method to obtain the solution. Another study, described in [179], proposes a resource allocation algorithm for two-tier cognitive HetNets with non-orthogonal multiple access (NOMA) technology. Specifically, this algorithm optimizes transmit power and subchannel allocation jointly, using both relaxation and worst-case approaches to convert the original problem into a convex one. In [180], authors investigate joint user association, bandwidth and power allocation, and caching deployment in heterogeneous fog radio access networks (Fog-RAN) by formulating an energy efficiency maximization problem as a non-convex problem, which is transformed into a consensus convex problem by relaxing binary variables. In particular, the resulting consensus problem can be efficiently solved using the alternating direction method of multipliers (ADMM). To address the sub-optimality of NOMA and orthogonal multiple access (OMA) in resource allocation, hybrid NOMA is applied in beyond 5G HetNets by the authors of [181]. They formulate a non-linear concave fractional programming problem to maximize energy efficiency, which is then transformed into a concave form using the Charnes-Cooper transformation. However, this transformation results in a mixed-integer non-linear programming problem. To solve it, the authors utilize a two-phase ϵ-optimal outer-approximation algorithm. In the multi-access enabled network [182], the authors study the transmission time allocation for NOMA-assisted computation offloading, aiming at minimizing the overall computing delay of all users. The problem is decomposed vertically into top and bottom problems, which are further solved by different optimization algorithms.

The mixed-integer nature of resource allocation problems, resulting from user association and subchannel allocation, poses a significant challenge to their solution. Pure optimization-based methods, as mentioned above, which are confined to specific scenarios, cannot be widely applied. Consequently, researchers have extensively explored the combination of optimization-based methods with other algorithms in HetNets. For instance, Qin et al. [183] used matching theory and a greedy-based algorithm along with Lagrangian multiplier method to solve the joint task offloading and resource allocation problem in HetNets. In their study [184], the authors explore the optimization of energy efficiency in NOMA HetNets with energy harvesting, where the subchannel allocation and power allocation have been resolved by modified Gale-Shapley matching algorithm and Lagrangian dual method, respectively. Moghimi et al. [185] employed alternative search method, Hungarian algorithm, and successive convex approximation to address the joint radio resource allocation and cooperative caching optimization problem in power-domain NOMA HetNets. To optimize the resource allocation in terrestrial-satellite HetNets, Zhang et al. [186] decouple the original highly non-convex problem into three subproblems: user association, bandwidth assignment, and power allocation. The authors propose a swapping-based algorithm to solve the user association subproblem, derive a closed-form expression for the bandwidth assignment, and use the Taylor expansion to approximate the power allocation subproblem, which is then solved iteratively. In addition, genetic algorithms have also been utilized in [56], [187] in conjunction with optimization-based methods to address the resource allocation problem in HetNets.

In [188], the authors investigate spectrum allocation and user association in HetNets using both mmWave and sub-6GHz frequency bands. They formulate cell load based spectrum allocation as a non-cooperative game to balance spectrum efficiency and reuse, proving the existence of a Nash equilibrium. For user association, they propose a distributed algorithm leveraging non-cooperative game theory, with the unique Nash equilibrium obtained through a fast converging Best Response algorithm. In [189], the authors propose a SDN-based HetNet to support efficient resource allocation for hybrid edge and cloud services. They establish an evolutionary game based service selection approach under incomplete information, and develop a Stackelberg differential game mechanism to trade cloud computing resources between different cloud and edge computing service providers. Extensive simulations validate the performance of the resource sharing mechanism and demonstrate convergence and equilibrium states. In [190], the authors utilize cooperative game theory to allocate beams in a massive MIMO HetNet in an energy-efficient manner, showing superior performance compared to existing approaches. The computation resource allocation problem in D2D relay-aided HetNets is tackled in [48], where the selfishness of users, which was ignored in previous works, is considered. To maximize utilities for both sellers and buyers, a group-based two-stage auction for relay-aided computation resource allocation (TARCO) is proposed in light of the computation offloading problem. Additionally, two modified algorithms based on TARCO are presented to achieve higher utility or total social welfare. Simulations illustrate the superior performance of these algorithms. In [191], considering the indispensable role of network slicing in improving isolation and flexibility, the authors propose a new network slicing architecture to tackle user association in ultra-dense HetNets. They formulate the problem as a one-to-many matching game and solve it through the UE-slice association algorithm, which enhances network performance as shown in numerical simulations.

AI/ML is regarded as an important feature of 6G, and has been widely used for solving various resource allocation problems. With the emergence of heterogeneous techniques in HetNets, radio utilization is expected to be improved by full-duplex TDD, where the TDD configuration is intractable for the high mobility HetNet. In light of this, the authors in [192] propose a DRL based online algorithm to dynamically allocate the resources, in which the deep neural network performs the complex information extraction and the dynamic Q-value iteration based RL achieve adaptively configuration. Different from the centralized resource allocation, the authors in [193] propose a distributed deep reinforcement learning based scheme for HetNets, where each BS allocates the resource only based on local information. To circumvent large state and action spaces resulting complicated network,

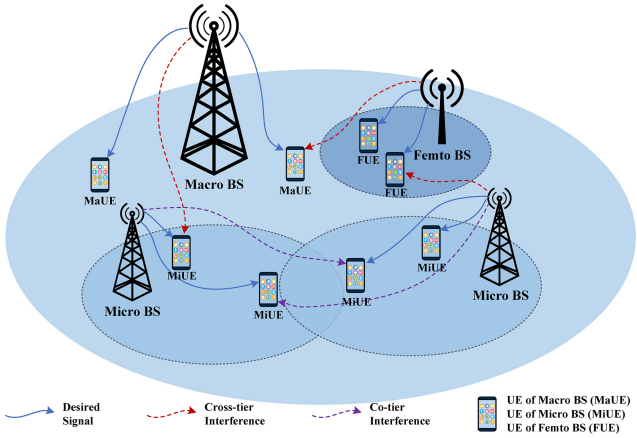

Fig. 14. Interference scenarios in HetNet.

a multi-agent Dueling Deep-Q Network (DDQN) is adopted, and the distributed and coordinated learning leads to a faster convergence than other learning algorithms. The extensive simulations also show the preferable performances of proposed algorithm compared to others. To cope with the high complexity from multi-RATs and serving multi-connectivity in future HetNets, the authors in [194] exploit a single-agent Deep-Q Network (DQN) algorithm and a multi-agent Deep Deterministic Policy Gradient (DDPG) algorithm to establish connections and allocate power, respectively. The single-agent at cloud-based edge servers interacts with multi-agents at RATs iteratively to learn the policy. The simulations show the superiority of this framework in terms of network utility and adaptability to dynamic changes. The authors investigate the dynamic user association and power control problem in HetNets with a tightly constrained computation time [195], where an unsupervised learning-based algorithm utilizing recurrent neural network (RNN) is proposed to resolve the problem. Comparing with the previous work [196], [197] using fully-connected-neural network (FCN) to achieve a reduced computation time, the proposed solution can adapt to the varying number of users as well as retain this advantage. Extensive simulations demonstrate a higher performance than traditional optimization-based methods.

2) Interference Management: The signals received at user or BS comprise desired signals, interference, and noise. Noise is typically caused by external factors such as electromagnetic interference, jitter, and distortion, which disrupt communication for all receivers. In contrast, interference refers to a signal that is undesired for some particular receivers but desired for others. In HetNets, interference can be categorized as either cross-tier or co-tier interference, as illustrated in Fig. 14. Co-tier interference arises from the same network tier, similar to interference in homogeneous networks, while cross-tier interference occurs when different network tiers share the same resources. As stated in [198], serious interference can significantly degrade network performance. Interference management in HetNets is even more complex and challenging than in homogeneous networks, given the trend of denser deployment [199], more heterogeneous RATs [27], and denser spectrum reuse [200], among others. In consideration of the aforementioned, we provide a summary of advanced interference management techniques, alongside a review of the state-of-the-art investigation on interference management.

Various interference management techniques have been investigated in recent literature, including mitigation, cancellation, and coordination. The first two techniques aim to alleviate interference within a single cell. For example, OFDMA employs orthogonal subcarriers in cells to mitigate interference, and interference rejection combining (IRC) suppresses interference in receivers by exploiting the difference between the desired and interference channels [201], and others. In terms of the coordination schemes, such as ICIC [63], enhanced ICIC (eICIC) [202], and CoMP are promising techniques to mitigate interference. Readers can consult the corresponding references for specific details regarding the techniques mentioned above. In the following subsection, we present a state-of-the-art survey on interference management in HetNets.

The authors in [203] proposed the successive interference cancellation and alignment (SICA) scheme to address the challenges posed by the unidirectional strong interference caused by in-building propagation loss and the diversity of transmit power, which have received limited attention in the literature for general K-user interference channels. SICA is designed to exploit interference rather than just aligning it, by transmitting the alignment streams and the superposition streams simultaneously, thereby improving the overall achievable rate and ensuring the overall degrees of freedom (DOF). To maximize the weighted sum rate, the joint transceiver designs for SICA are formulated and solved alternately by solving a sequence of equivalent weighted mean square error problems. In [204], a scheme is proposed that combines interference avoidance and CoMP techniques in terms of SINRs in machine-type HetNets. Specifically, for outdoor machine-type devices whose SINRs are lower than a threshold, one of the techniques in the scheme is employed, namely, SBSs transmit nothing to reduce interference or cooperative data signals to collaboration. To eliminate interference between users in NOMA-enabled HetNets that share spectrum between users of OFDM and orthogonal time-frequency division (OTFS), the authors propose an iterative multiuser detection algorithm based on successive interference cancellation (SIC) [205]. Specifically, the receivers incorporate two detectors: one for low-mobility users with OFDM and one for high-mobility users with OTFS. They exchange their output decisions and use them in the SIC process. Extensive simulations and extrinsic information transfer chart analysis demonstrate the superiority of the proposed algorithm in terms of convergence rate and error performance. The authors propose the optimal SIC ordering and power allocation (JSPA) algorithm to maximize the sum-rate in NOMA networks [206]. The decoding order of users is dynamically updated during power control. However, due to its forbidding complexity, the authors adopt a suboptimal scheme with fixed SIC ordering. Specifically, they investigate a lower-complexity near-optimal strategy to jointly optimize rate and power. Furthermore, a semi-centralized framework is proposed for larger-scale two-tier HetNets to manage interference efficiently. The adoption

TABLE VIII
LITERATURE ON POWER ALLOCATION AND ENERGY EFFICIENCY OPTIMIZATION IN CF NETWORK

Year	Ref. No.	Scenario	Main Discovery
2015	[208]	downlink	Showed that the power optimization problem of non-convex nature transforms into a geometric programming problem, which is solved by an iterative algorithm.
2019	[209]	downlink	Proposed sequential algorithm for GEE maximization to solve the non-convex problem of power distribution for maximum global energy efficiency.
2019	[210]	uplink	Showed that SCA and heuristic suboptimal solutions transform the power allocation problem into a standard geometric programming (GP) problem and iteratively solve the decoupled subproblems.
2020	[211]	downlink	Proposed the nonlinear power control of SDP-optimized downlink transmission.
2020	[212]	uplink & downlink	Showed the iterative process of solving ZF-based problems through the ICA framework and Dinkelbach method so that only a simple convex program is solved for each iteration.
2020	[213]	uplink	Showed that the successive internal approximation (SIA) technique transforms the max-min quality of service (QoS) power control problems into a series of convex optimization problems that can be solved iteratively.
2020	[214]	downlink	Proposed the efficient power optimization algorithm based on the Lagrangian multiplier method to improve the performance of the incoherent and coherent joint transmission.
2020	[215]	downlink	Showed that it solves the mixed integer second-order cone program to obtain the optimal global solution of the non-convex problem.
2020	[216]	uplink & downlink	Showed that jointly optimized uplink and downlink power control coefficients minimize the total transmit energy consumption while satisfying target SINRs.
2020	[217]	downlink	Proposed the deep neural network (DNN) to solve the non-deterministic polynomial (NP) problem of promptly maximizing each user's minimum capacity.
2021	[218]	uplink	Proposed the access point (AP) allocation algorithm to improve decentralized schemes' spectrum and energy efficiency.
2021	[219]	uplink & downlink	Proposed the alternating optimization algorithm to solve the problem of highly coupled non-convex optimization.
2021	[220]	uplink	Proposed two GP-based successive approximation algorithms for maximizing total spectral efficiency and maximizing total energy efficiency.
2021	[221]	uplink & downlink	Proposed the first-order algorithm to find the initial feasible point and approximate optimal solution of the non-convex optimization problem with maximum energy efficiency.
2021	[222]	uplink	Showed the maximum-minimum fair power optimization algorithm by solving geometric programming problems, ensuring that users receive consistent and good service in any geographical location.
2022	[223]	downlink	Showed that a new centralized framework is adopted to solve a series of simpler power distribution subproblems by incorporating fractional programming, non-cooperative game theory, and gradient-assisted binary search (GABS) algorithms.
2022	[224]	uplink & downlink	Proposed the iterative power control algorithm based on the framework of the accelerated projection gradient (APG) method.
2022	[225]	uplink	Proposed two new energy delay sensing power control strategies to minimize energy delay costs online by employing Bernoulli-Bandit learning (BBL) and Gauss-Bandit learning (GBL).
2022	[226]	downlink	Proposed the new block-QT technique to optimize the global energy efficiency (GEE) of the non-convex network center and decompose the GEE optimization into more specific convex problems.
2022	[227]	uplink & downlink	Showed that maximizing GM-rate solves power allocation problems and perceived quality of service (QoS) network energy efficiency problems.
2022	[228]	downlink	Proposed the suboptimal algorithm with high computational efficiency based on SCA to solve optimal power control's maximum and minimum problem.
2022	[229]	downlink	Proposed the maximum-minimum power control strategy based on the accelerated projection gradient (APG) method.

of a multi-tier architecture with lower power BS and smaller cells that share the same spectrum with multiple users leads to severe interference in massive MIMO HetNets, resulting in significant degradation of network performance. To this end, the authors propose a feedback mechanism on an existing algorithm and utilize evolutionary game theory to effectively reduce interference [207].

D. Resource Allocation in CF Network

Transmission power allocation and global energy efficiency optimization in cell-free networks are essential ways to optimize the system's performance. By adjusting the transmission power of each AP, transmission power allocation aims to maximize system capacity. To improve the system's performance wholly, global energy efficiency optimization is to find the best balance between the system's spectral efficiency and energy efficiency. Tab. VIII summarizes the literature for transmit power allocation and global energy efficiency optimization. Through the joint modeling of energy efficiency and spectral efficiency, the problem was the non-convex power distribution problem by maximizing global energy efficiency.

For downlink of large-scale cell-free MIMO, the Sequential Algorithm for GEE Maximization (SAGM) was applied in [209] for cell-free and user-centric large-scale MIMO scenarios at millimeter-wave frequency band. This algorithm solved the non-convex power distribution problem by maximizing the global energy efficiency of downlink. It proposed a low complexity power distribution rule aiming at maximizing global energy efficiency. The analysis showed that this allocation algorithm can effectively increase the system's energy efficiency and average rate per user.

Max-Min Fair algorithm (MMF) combined with LSFD and power control can maximize the minimum SE among UEs. Due to the highly coupled non-convex problem, [219]

proposed an alternating optimization algorithm. It combined the closed LSFD vector with binary-searching, and can ensure the algorithm's convergence while finding the optimal global solution.

Non-orthogonal multicast and unicast transmission resource allocation for massive cell-free MIMO was a non-convex optimization problem with complex constraints. For solving the problem, authors in [221] proposed a first-order algorithm (FOA) to find the initial feasible point and approximate optimal solution. It was much less complex than general second-order algorithms, and can obtain almost the same performance.

Mixed integer second-order cone programming method can solve non-convex optimization problems, but the computational complexity is high for real-time implementation. Therefore, [215] improved it and proposed two suboptimal algorithms with lower complexity. The algorithm used the inherent group sparsity and optimized transmit power solution in the problem, respectively. The computational complexity was much less than the branch delimitation method. The power consumption was only about 20% higher than the global optimal method, which can be applied into large cell-free networks.

In the uplink of cell-free massive MIMO network, authors in [210] decoupled the energy efficiency maximization problem into two sub-problems. One was the receiver filter coefficient design and other is power allocation. The receiver filter coefficient design was formulated as a generalized eigenvalue problem. The power allocation problem was transformed into a standard geometric programming (GP) problem through successive convex approximation (SCA) and heuristic sub-optimal schemes. The numerical results showed that the proposed algorithm can achieve almost twice the total energy efficiency than that of equal power allocation method. In [228], it modeled power allocation control as a max-min problem and proposed a sub-optimal algorithm with high computational efficiency based on SCA.

In order to obtain higher spectrum efficiency and energy efficiency in the in-band full-duplex (FD) CF mMIMO wireless network, two low-complexity transmission methods based on ZF were proposed in [212]. The iterative process of solving ZF-based problems was derived through the ICA framework and Dinkelbach method. Then, a simple convex programming problem can be solved within each iteration. It was worth noting that the proposed ZF-based transmission method consumed less time than a simple method with MRC.

In Rayleigh fading channel of the uplink transmission, authors in [213] considered max-min QoS power control problem in generalized cell-free massive MIMO systems. The successive internal approximation was used to transform the NP complex problems of related optimization problems into a series of convex optimization problems, which can be solved iteratively with low complexity.

In [226], a new block-quadratic transformation (block-QT) scheme was proposed to optimize the global energy efficiency (GEE) of the center of a non-convex network. This scheme combined block optimization and quadratic transformation methods to decompose GEE optimization into simpler convex sub-problems. For non-convex optimization decomposition, simulation results demonstrated that the Block-QT scheme was better than the SCA method.

E. Lessons Learned

From the review of resource virtualization and allocation in various RANs, we can learn the following lessons so as to provide guidance to the development and design of a different RAN for 6G:

- It is suggested that resources should not be tightly bound with BSs or cells, mobile operators, and specific usage (i.e., uplink or downlink). In other words, spectrum can be utilized in an ultra flexible manner. In this way, more resources can be aggregated for serving the users, even if the total resources do not increase. Then, resource allocation can potentially become more powerful, because more available resources lead to larger solution space for resource allocation. It is also helpful to aggregate resources from a larger network consisting of more heterogeneous BSs and resources. Furthermore, by exploiting virtualization techniques, it is equivalent to have more resources to be utilized, owing to statistical multiplexing gain.
- It is further suggested that a real-time RAN intelligent controller (RIC) should be deployed at the edge for realizing efficient and adaptable resource scheduling. Under the large-scale and heterogeneous network scenario, resource scheduling will also become more complicated to implement in practice, while the performance improvement gained from resource scheduling can only be achieved when it is feasible. Also, resource allocation has different problem formulations in different networks and under different conditions, and it is very hard to find the universal solution for resource scheduling problems with traditional optimization methods. Thus, AI-based resource scheduling should be further investigated and deployed at the real-time RIC, with enhanced user intention sensing, control, and computing resources support.

VI. FD-RAN: Enabled Technologies and Future Directions

This section first presents several studies [74], [230], [231] enabled by the FD-RAN architecture, demonstrating the advantages of FD-RAN. Due to page limit, we only briefly introduce the background, proposed method, simulation environment and results of each study. More information can be found in the original published papers. Then, the future directions of FD-RAN are discussed, including the potential applications of FD-RAN for both general scenarios and emerging services of 6G, as well as the open issues and potential research directions.

A. Cooperative Uplink Reception

The uplink network consists of densely deployed UL-BSs and is facilitated by the edge cloud, which serves as the coordinator for all UL-BSs and users. Users are served by

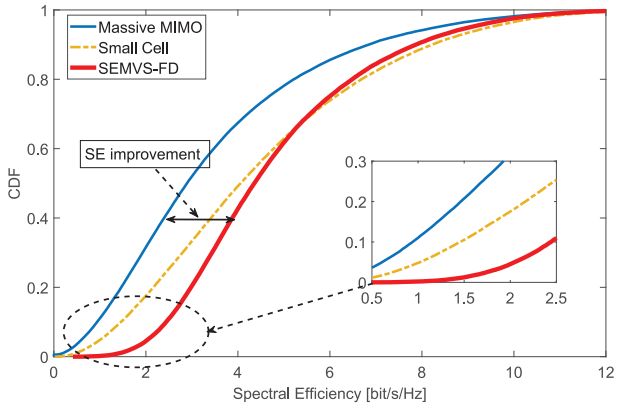

Fig. 15. The CDF of users' spectral efficiency for different networks.

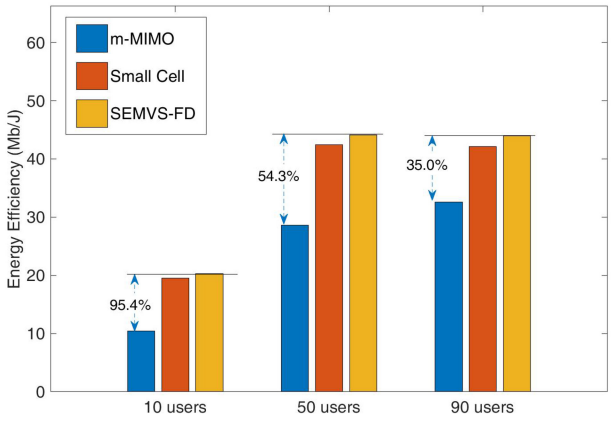

Fig. 16. Energy efficiency of users for different networks.

multiple cooperative UL-BSs to enhance the signal reception performance. The network must determine the optimal scheduling of UL-BSs for each user, considering factors such as fronthaul capacity constraints and user fairness.

In [74], the proposed resilient cooperative reception framework adopts a two-tier signal combination approach to address large-scale fading, which consists of localized signal combination at the base stations and centralized signal combination at the edge cloud. Data signals from users are transmitted to cooperative virtual service clusters (VSCs) of UL-BSs, where the received signals from multiple antennas are combined and forwarded to the edge cloud for further processing. The channel statistical information is collected and periodically transmitted to the edge cloud, providing global channel information for centralized signal combination by the edge cloud. Additionally, transmit power control is utilized to reduce interference between users and improve network spectral efficiency, as well as to reduce the overall power consumption of UEs compared to transmitting at maximum power.

In the simulation, the following scenario is considered: one control BS, $M = 64$ UL-BSs with antennas $N = 4$, and $K = 30$ users distributed in a $1000m \times 1000m$ square area. All the UL-BSs and users are independently and uniformly distributed in the area of interest. The distance-based wrap-around simulation strategy is adopted, and the performance of the network with 30 active users/km^2 and 256 antennas/km^2 can be acquired. In the area, all UL-BSs receive both signal and interference from all directions simultaneously. The massive MIMO based cellular network is also evaluated for comparison, keeping the same number of antennas with $M_c = 4$ BSs and $N_c = 64$ antennas in the same area. The small cell scenario is also considered, in which $M_s = 64$ small cells, each with $N_s = 4$ antennas, are uniformly distributed in the same area. The channel follows spatially correlated Rayleigh fading at 2GHz carrier frequency, where the large-scale fading is based on the 3GPP Standards [232] and the spatial correlation matrices is generated employing [233]. The transmit power is up to 200mW with a bandwidth 20MHz, and the noise power spectrum density is -174dBm/Hz with a noise figure 7dB.

As shown in Fig. 15, the SE performance of the FD-RAN system with the spectrum-efficiency maximized virtual service cluster selection (SEMVS) algorithm is generally superior to that of the massive MIMO cellular network and the small cell network, particularly for users experiencing poor channel conditions. This improvement is primarily due to the cooperative reception provided by multiple UL-BSs in FD-RAN. Additionally, the distance between users and base stations is smaller in FD-RAN compared to the massive MIMO cellular network, leading to a stronger received signal strength at UL-BSs. Fig. 16 further illustrates the comparison of EE between the different networks. As a result of the aforementioned factors, the FD-RAN uplink demonstrates superior EE.

B. Cooperative Downlink Transmission

In FD-RAN, downlink transmission is fully separated from uplink transmission, enabling flexible deployment of DL-BSs based on UEs' traffic distribution and downlink coverage needs. Additionally, a UE can be served by multiple DL-BSs through cooperation. Specifically, coordinated beamforming is employed, where multiple BSs act as one virtual BS and a UE can choose a subset of the total antennas in this virtual BS. Data is transmitted to the UE by these antennas and BSs simultaneously. However, acquiring CSI is a major challenge for downlink coordinated beamforming in FD-RAN, as it cannot be directly fed back to the DL-BS. Also, channel reciprocity cannot be utilized as there is no duplex channel for data transmission in FD-RAN.

In FD-RAN, the radio antennas for both uplink and downlink are physically distinct. Hence, a general channel estimation and feedback process is necessary for downlink association, link adaptation, beamforming, and other operations [230]. The downlink problem involves the association of multiple BSs and UEs and the dynamic allocation of resources, with the aim of maximizing the weighted sum rate. Fig. 17 illustrates the coordinated beamforming approach with two-stage channel estimation error, as described in [234]. The channel estimation error between downlink pilot transmission and uplink feedback is modeled as a mixed-integer non-convex programming problem, which can then be decomposed into multi-connectivity and beamforming subproblems.

The simulation focuses on a 7-hexagon urban micro cells, each having a radius of 200m. Each cell consists of three

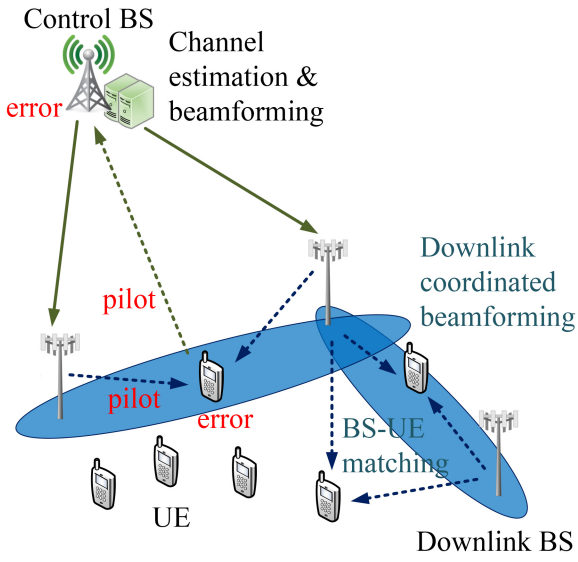

Fig. 17. The system model of downlink coordinated beamforming.

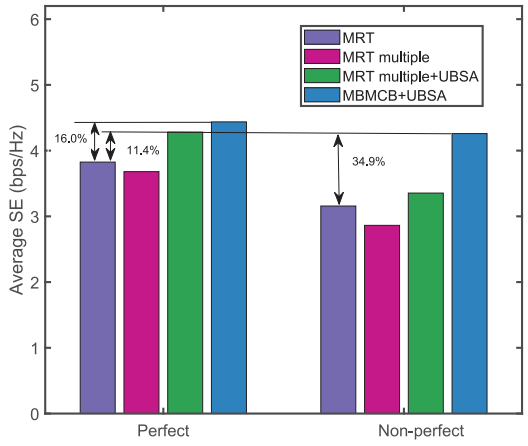

Fig. 18. Average SE versus different channel conditions.

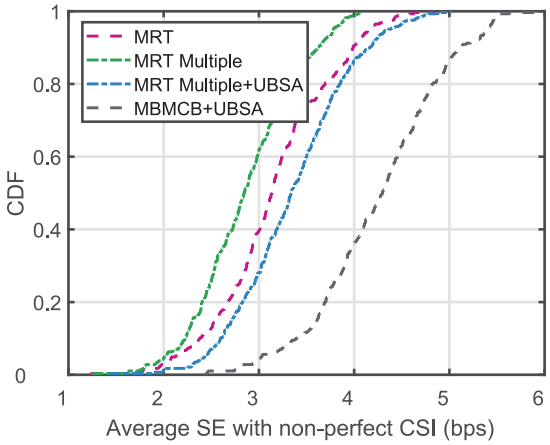

Fig. 19. Average SE versus different channel conditions.

sectors (i.e., three BSs), and is equipped with a linear antenna array with $N = 10$ antennas. There are three UEs generated independently in each cell. A circular exclusion region with radius of 20m is considered. The maximum DL transmit power is set as 43 dBm. The thermal noise figure of each UE is set to -114 dBm. The communication bandwidth is set to unit 1. The maximum number of BSs that a UE can connected to is restricted to 3. The importance weights δ_k for the proportional fairness is set as 1 for each UE. In addition, the DL channel is generated according to standard 3GPP 36.873 urban micro (UMI) channel using the QuaDRiGa simulation environment. The performance of following algorithms are compared:

- MRT: MRT beamforming is a traditional linear beamforming method with low complexity. Here, the single-connectivity MRT (i.e., each UE access to only one BS) is used as a baseline algorithm.
- MRT Multiple: This algorithm considers the MRT beamforming in terms of the multi-connectivity. Besides, this multi-connectivity is simply depending on the DL RSRP.
- MRT Multiple+UBSA: This algorithm adds the UE-BS swap-matching algorithm (UBSA) on the basis of multi-connectivity compared to MRT Multiple algorithm.
- MBMCB+UBSA: This algorithm includes multi-BS multi-UE coordinated beamforming (MBMCB) and UBSA algorithm.

Fig. 18 presents the comparison of the sum capacity obtained by different algorithms under both perfect and non-perfect CSI conditions. The MBMCB+UBSA method shows an average SE improvement of 16.0% under the perfect channel condition and a 34.9% improvement under the non-perfect channel condition. It can be seen that MBMCB+UBSA still offers an average SE improvement of nearly 11.4% under the non-perfect CSI condition compared to its performance under perfect CSI.

The CDFs of the non-perfect CSI conditions for different algorithms are shown in Fig. 19. The results demonstrate that simple multi-connectivity without association scheduling or power optimization does not result in any SE gain. In comparison, the MRT+UBSA method only provides a slight SE improvement, whereas the MAMCB+UBSA method significantly increases the average SE. This result highlights the effectiveness of the proposed MAMCB beamforming method.

C. Flexible Resource Allocation

The rationale behind this study is to demonstrate the flexible spectrum utilization pattern of FD-RAN, which is different from TDD and FDD in traditional RANs. In FD-RAN, the whole spectrum can be used for either uplink or downlink, depending on the traffic demands. This is owning to the physical separation of uplink and downlink networks. As a result, the spectrum utilization efficiency in practical scenarios can be improved.

In FDD, dedicated frequency bands are allocated for uplink and downlink transmission. However, the method of spectrum usage in FDD is not flexible or efficient enough to accommodate the changing user traffic requirements. Additionally, FDD requires frequency bands to be assigned in pairs, with the same bandwidth for both uplink and downlink, resulting in a rigid and less efficient use of spectrum resources that can lead to underutilization of uplink and downlink capacities and impose limitations on frequency band planning. Although

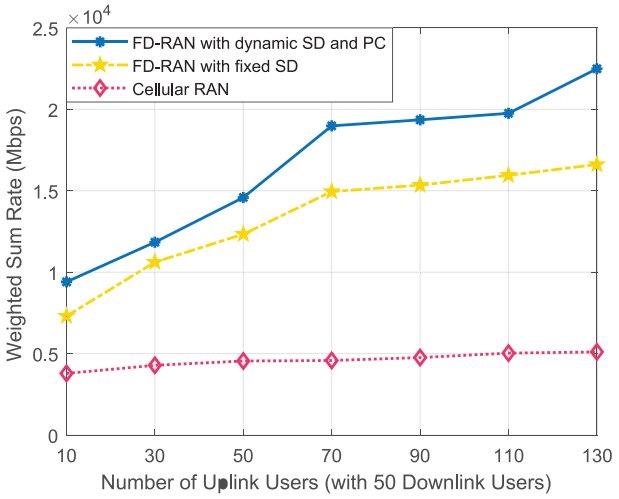

Fig. 20. Weighted sum rate comparisons under different user load.

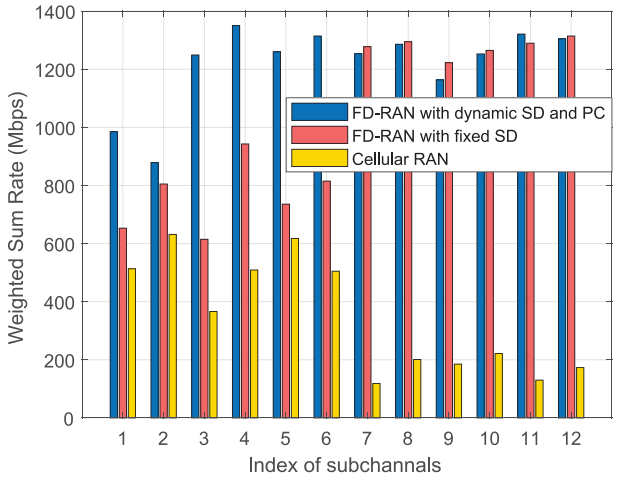

Fig. 21. Weighted sum rate comparisons of each subchannel through flexible resource allocation strategies.

TDD avoids the issues faced by FDD, it still lacks flexibility in practical deployments and cannot adapt to varying traffic demands.

The proposed flexible spectrum allocation method for FD-RAN aims to enhance the utilization efficiency of the available spectrum [231]. Unlike the traditional FDD and TDD modes, which are designed for single cells, FD-RAN is similar to C-RAN and operates with centralized control over a larger region. The new method allows the entire available spectrum to be used for both uplink and downlink transmission and adapts the allocation of frequency bands to respond to changes in user service requirements, providing improved efficiency in spectrum utilization.

Overall, the dynamic spectrum allocation method in FD-RAN operates as follows. The centralized controller located at the edge cloud collects the uplink and downlink service requirements of all users in the local area. Based on these requirements, the controller dynamically divides the available spectrum between uplink and downlink transmission. To do so, it selects appropriate subchannels and uplink and downlink base stations for each user, taking into account factors such as the fronthaul capacity of the base stations, user QoS, and the system's overall capacity.

The simulation considers a dense urban area consisting of 7 hexagonal virtual cells, where each virtual cell has a radius of 100 m and contains 1 DBS in the center and 5 UBSs that are uniformly distributed. In the FD-RAN, each BS has 3 sectors facing different directions and there are 105 UBS sectors and 21 DBS sectors in total, where each UBS sector has 2 receiving antennas and each DBS sector has 10 transmitting antennas. Each sector can be regarded as an independent BS for uplink and downlink transmissions. QuaDRiGa is used to generate the transmission channel based on the 3GPP Standards. The network has a total bandwidth of 120 MHz and 12 subchannels, each having a bandwidth of 10 MHz. The performance of three resource allocation methods is compared and evaluated through simulation results presented in Fig. 20 and Fig. 21. The evaluated methods are the FD-RAN with spectrum division (SD) and power control (PC), the FD-RAN without SD, and the traditional cellular RAN.

In Fig. 20, the performance of three different resource allocation methods is evaluated and compared under varying user loads. The number of uplink users is varied from 10 to 130, while the number of downlink users is kept constant at 50. Here, a saturated traffic model is assumed such that each user always has data to transmit if allocated resources. The results show that the FD-RAN with fixed SD outperforms the traditional cellular RAN, with a weighted sum rate that is on average 1.45 times higher. Additionally, the FD-RAN with dynamic SD and PC further improves the performance, with a weighted sum rate that is on average 30% higher than that of the FD-RAN with fixed SD. These results demonstrate the superiority of the FD-RAN architecture and the effectiveness of the proposed resource allocation algorithm.

In Fig. 21, the number of uplink and downlink users are both 50. It is observed that the weighted sum rate per subchannel in the FD-RAN schemes surpasses that of the traditional cellular RAN. Additionally, a comparison between the FD-RAN without SD scheme (with subchannels 1-6 allocated to downlink and subchannels 7-12 allocated to uplink) and the FD-RAN with dynamic SD and PC scheme (with subchannels 1-2 allocated to downlink and subchannels 3-12 allocated to uplink) demonstrates the superior spectrum utilization and increased system capacity achieved by the latter.

D. Lessons Learned

From the review of above studies, we learn lessons on how FD-RAN enables these PHY and MAC technologies with its two key features.

- Multi-BS cooperation is enabled by the physical separation of control and data BSs. In both uplink and downlink, the UE is actually served by a dynamic virtual cooperation set of BSs. In order to support multi-connectivity for UEs as a default mode, control plane needs to be simplified, so that the association between UE and multiple UL-BSs/DL-BSs can be flexibly changed.

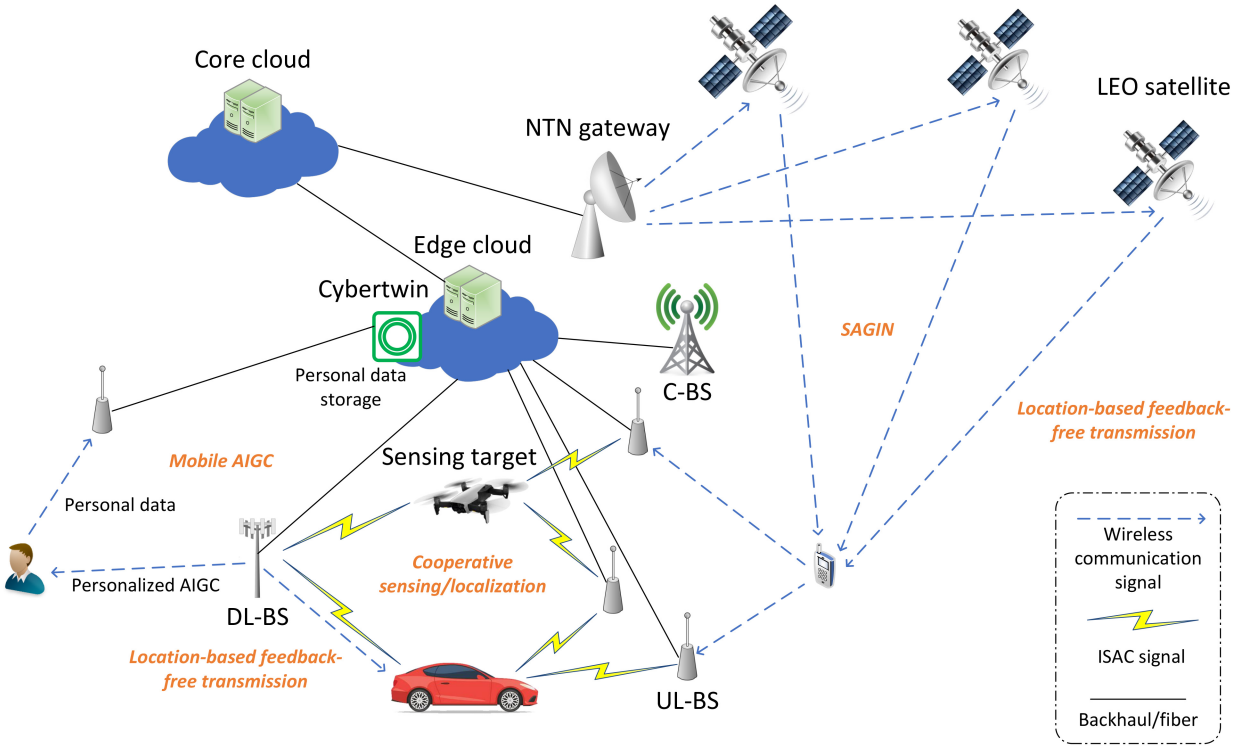

Fig. 22. Integrating FD-RAN with emerging services.

In FD-RAN, since the C-BS takes over all the control plane communications with UEs, the data BSs and UEs do not have to setup and maintain RRC connections to transmit data, reducing the overheads greatly.

- Flexible spectrum utilization is enabled by the physical separation of uplink and downlink networks. In the current 5G network, spectrum utilization for uplink and downlink is static. The rationale behind is simple, because in traditional RAN architectures, uplink and downlink are tightly coupled to serve the users' demands together. However, in FD-RAN, the uplink and downlink networks become independent, thus they do not need to consider each other and can only focus on providing the uplink or downlink services with the available spectrum, which can be flexibly adjusted according to the actual traffic demands.

E. Applications of FD-RAN for 6G Emerging Services

The application scenarios of FD-RAN can be both general and specific. FD-RAN can serve as the supplement of existing 5G networks, owing to the decoupling of uplink and downlink BSs. For example, independent DL-BSs can be deployed where downlink traffic demands are high. In non-public network or vertical scenarios, deploying FD-RAN in standalone mode is also more advantageous than deploying traditional RAN, since the UL-BSs and DL-BSs can be deployed more flexibly based on the traffic demands.

FD-RAN can also be applied for several emerging services of 6G. Specifically, it can be combined with integrated sensing and communications (ISAC), space-air-ground integrated network (SAGIN), and AI generated contents (AIGC), respectively, as illustrated by Fig. 22. Although FD-RAN is designed separately, when integrated with these services, we find it can show superiority in some aspects.

1) FD-RAN and ISAC: Due to lack of frequency resources, 5G is moving to higher frequency bands located at tens of GHz, e.g., 28 GHz and 60 GHz. Such mmWave bands have also been used for traditional radar sensing. Since the sensing, especially localization capability is regarded as a significant service of 6G, how to integrate sensing and communication in the same system has become a major research direction for 6G [235]. The ISAC system can utilize signals for both data transmission and sensing, thereby saving the precious spectrum resources. Thus, the waveform design for transmitting such ISAC signals becomes critical [236].

According to [237], ISAC can be classified into device-based and device-free. The former is an active sensing process that incorporates signal transmission and/or reception of the sensing targets, while the latter is a passive sensing process that requires no additional capabilities from the sensing targets. We focus on the device-based ISAC, which can be further classified based on whether the sensing/localization is cooperative or not. Cooperative localization is more powerful since more information can be provided. However, it is also more challenging since the sensing signals from different sources need to be combined. The network itself also needs to support the cooperative transmission, such as relay and device-to-device communications mentioned in [237].

It is straightforward for FD-RAN to perform cooperative localization, due to its uplink-downlink decoupled architecture. Specifically, in cooperative sensing, the sensing signals can be

transmitted from the DL-BS to the UE, and are then received by the UL-BSs, as illustrated in Fig. 22. Thus, there is no need to consider the physical separation of signal transmission and reception, since it is already realized in FD-RAN. The signal combination and processing can be done at the edge cloud, and the location information of UE can be delivered to the C-BS. Furthermore, C-BS can utilize UE's location to perform location-based feedback-free transmission directly. Since the location feedback process is saved, the performance of FD-RAN can also be improved. To conclude, combining ISAC and FD-RAN is beneficial to each other and is therefore a very promising solution.

2) FD-RAN and SAGIN: The satellite and aerial-based communications are indispensable for providing ubiquitous connectivity, which is an important objective for 6G. In SAGIN, different layers of networks are integrated, including the GEO, MEO, and LEO satellites as well as high-altitude platforms (HAPs) [238]. 3GPP has also defined the architectures of non-terrestrial networks (NTNs) [239], in which the air interface of an NG-RAN is replaced by the NTN gateway and satellite. Thus, the NTNs are totally transparent to the mobile users.

Specifically, we consider the ultra-dense LEO satellite-terrestrial integrated 6G [240]. LEO satellites can have much lower latency and higher throughput than GEO/MEO satellites, and can provide more coverage and reliability than HAPs. These advantages make the ultra-dense LEO satellite network (e.g., Starlink) a competitive substitute for the terrestrial mobile networks. Besides, in the future cloud-based Internet, most services are deployed in the cloud data centers. Then, the NTN gateways can be directly connected with the cloud, such that mobile users are only one-hop away from the services.

However, since the transmit power of mobile devices is low, the uplink of traditional satellite-based NTN is not efficient and may cause high energy consumption of UEs. To this end, a natural choice is to physically decouple the uplink and downlink, which requires a new fully-decoupled architecture, namely FD-RAN. The downlink transmission can still be carried out by the LEO satellites, and the location-based feedback-free transmission mechanism used in FD-RAN can be utilized. On the other hand, the UL-BSs in FD-RAN can be used for receiving uplink signals from UEs. Cooperative transmission techniques can also be adopted for both uplink and downlink. Therefore, under the FD-RAN architecture, the advantages of ultra-dense LEO satellite networks can be fully exploited, while the disadvantages can be avoided, making the FD-RAN a promising architecture for the future 6G SAGIN.

3) FD-RAN and mobile AIGC: AIGC refers to the contents that are generated by AI models instead of human users, such as the large language models (e.g., GPT-4 [241]) used for generating texts, and the diffusion models used for generating images (e.g., Stable Diffusion [242]). Since AIGC can greatly improve the producing efficiency, it is regarded as the most important service for the future Internet. Therefore, it is also necessary to consider how to provide AIGC services in the mobile network.

According to a recent survey on the mobile AIGC network [17], integrating AIGC services and mobile networks has several advantages. On the one hand, the models can be deployed at the edge cloud, such that users can access the services with lower latency. On the other hand, personalized services can be provided through fine-tuning models based on users' own data and requirement. For the purpose of privacy, the fine-tuned models can be deployed on the local edge servers.

FD-RAN can further enhance the mobile AIGC network in several aspects. In FD-RAN, Cybertwin serves as each user's personal data keeper at the edge cloud. Thus, the personalized model can be fine-tuned through the private data kept by Cybertwin, and there is no risk of leaking the data to third-party. Furthermore, at the RAN side, FD-RAN provides personalized transmission service through user-centric resource allocation. Although AIGC services are resource-consuming, stronger service quality guarantee can be achieved with FD-RAN. Moreover, as shown in the lifecycle of AIGC services in [17], pre-training the large model requires huge amount of data, which need to be collected through the mobile network in more and more situations, such as industrial IoT. With the help of uplink and downlink decoupling in FD-RAN, the huge traffic of data collection can be satisfied through deploying more UL-BSs, and the infrastructure and energy costs can be greatly reduced.

F. Open Issues

In this subsection, we discuss some open issues in FD-RAN and point out some future directions.

1) Delayed/No Feedback: Owing to the decoupling paradigm of FD-RAN, the feedback mechanism in traditional RAN is weakened. Although the C-BS can be utilized for feedback, additional delay will be incurred, leading to inaccuracy or untimeliness of feedback information. Specifically, in FD-RAN, delayed feedback will have negative impacts on CSI and HARQ. Potential research directions on these two aspects will be discussed in the following, respectively.

In state-of-the-art MIMO techniques such as spatial multiplexing, channel information is required for precoding. In TDD systems, CSI can be easily acquired through sending pilots for channel estimation and utilizing channel reciprocal. In FDD systems, CSI is usually feedback in a compact or quantized format so as to reduce overheads. For either systems, timeliness is important since channel state varies fast. However, this means more resource consumption. In FD-RAN, since there is no direct feedback, the problem becomes even worse. To this end, a novel method, namely no-feedback MIMO transmission is proposed [243]. The channel information is solely acquired from UE's geolocation, with the help of AI and historical channel data, which are highly correlated with location. Although no-feedback MIMO transmission will lose some performance due to lack of realtime CSI, it can save the resource consumption of pilots and feedback overheads, and is less sensitive to feedback delay.

HARQ plays an important role in error control at the MAC layer. To deal with the latency in HARQ process, two possible approaches can be considered. On the one hand, HARQ in

non-terrestrial networks has been studied [244]. 3GPP Release 17 also develops solutions for HARQ with large feedback latency. These solutions can be exploited by FD-RAN, with the C-BS used for HARQ feedback. On the other hand, the HARQ mechanism can also be deactivated. Instead, error control can be achieved at higher layer, e.g., transport layer. To reduce packet retransmission delay, the transport layer can be implemented inside the RAN, namely stacked onto the end to end connection between UE and the user plane function at core network. Furthermore, QUIC protocol can be adopted for better transport layer performance.

2) Massive Low-latency Control Signalling: In FD-RAN, the control signaling happens between UEs and the C-BS. The C-BS, which uses low-frequency bands for a wide coverage area, may face a large number of UEs accessing it, each requiring frequent transmission and receipt of control signals. Furthermore, these control signals must meet stringent requirements, such as sub-1ms latency and 99.9999% reliability. The traffic pattern of control messages must also be taken into account, including the mix of small, periodic control packets and signaling bursts, as well as different priorities of different control messages. In conclusion, the C-BS must satisfy both the uRLLC and the massive machine type communication (mMTC) [245] requirements in 5G, which is a challenging task. Thus, it is imperative to employ appropriate multiple access schemes to achieve this goal.

To meet the requirements of both URLLC and mMTC in 5G, the 3GPP has proposed the grant-free random access (GFRA) mechanism [246] since Release 15. GFRA replaces the traditional four-step grant-based request-triggered transmission with a more efficient two-step procedure, in which the UE can transmit small packets along with necessary control information [247], resulting in reduced uplink transmission latency. Reliability is improved by implementing advanced hybrid automatic repeat request techniques. GFRA can either pre-allocate frequency resources or use contention for frequency sharing, with the former being suitable for periodic transmission and the latter for sporadic packets. However, when the number of accessed UEs increases, contention-based access may experience conflicts and reserved frequency resources may not be sufficient, leading to reduced quality of service. In such cases, other access control mechanisms such as access class barring and back-off may be implemented, especially considering the different priorities of the packets.

For downlink support in FD-RAN to address the requirements of mMTC and URLLC, a multiple access mechanism that can accommodate the simultaneous transmissions from a large number of UEs is required. This demands the slicing of wireless resources on additional dimensions, such as space, power, and code. In this context, NOMA [248], [249] can prove to be a promising solution. In NOMA, non-orthogonal utilization of resources can be achieved in the power or code domain, enabling multiple users to be served by the same time-frequency resource block. For instance, in power domain NOMA, signals from different users are superposed on the same resource block and differentiated through different power levels. The receiver then performs SIC to decode the signals. Additionally, NOMA can be integrated with other technologies like MIMO to make better use of the spatial domain and support massive connectivity. 3GPP has also considered NOMA as a potential solution within the framework of multiuser superposition transmission (MUST) [250].

3) Cybertwin-Assisted Personalized Service Provision: In 6G, the focus should be not only on improving system-level performance metrics such as network throughput, but also on better serving its users. To accomplish this, there must be a way to understand and characterize a user's service expectations. However, traditional QoE metrics do not consider the subjective opinions of individual users. There also needs to be a communication channel established between users and the network, allowing the network to be aware of each user's personalized requirements. Lastly, the network should implement a resource allocation scheme that takes into account each user's personalized requirements and their respective priorities, as the total network resources are limited.

In FD-RAN, Cybertwin [70] serves as the user's agent, acting as the user's point of entry to the Internet. It is responsible for communicating the user's requirements to the network. Prior to using the network, the user should configure Cybertwin with his or her desired performance metrics for each service, such as data rate and latency, along with a value associated with the service. This allows the network to compare and prioritize the relative importance of different services for different users. Dynamic pricing [251], [252] of network resources is also used to reflect the real-time balance between demand and supply. When the price of resources is high, the network must prioritize services with higher values and the corresponding users must pay a higher price. The network operator may sign SLAs with different prices for each user or a novel utility function considering both transmission performance and user value can be utilized to allocate resources.

4) Energy Consumption of UEs and BSs: It is imperative to reduce the energy consumption of both UEs and BSs for 6G. Due to the decoupling paradigm, FD-RAN has the potential to optimize the energy usage. From the UE's point of view, its transmit power needs to be reduced. In FD-RAN, since the UL-BSs and DL-BSs are physically decoupled, the UE can be associated with the nearby UL-BSs so as to reduce the uplink transmission distance. Furthermore, uplink receive diversity, signal combining and resource cooperation techniques can be adopted in FD-RAN. In both ways, the UE can use less power to transmit data. For the control plane, energy-efficient communication protocols can be developed so as to minimize signaling overhead.

To reduce the power consumption of BSs, one of the most efficient ways is BS sleeping. Thus, we should consider from the network's perspective: when the total traffic demands are low, some of the BSs can be deactivated. In FD-RAN, this can be achieved with C-BS performing central control. With C-BS, UEs can at least be connected by the control channel, such that coverage holes can be avoided. Also, the solution space for optimizing total energy consumption becomes much larger, since the UL-BSs and DL-BSs are separated. Then, it is worth studying an efficient strategy to balance the energy costs and users' QoS.

VII. Conclusion

In this survey, we have presented an evolutionary perspective on the 6G RAN architectures. At such a time when many 6G initiatives are being launched or are about to be launched, the meaning of this survey is significant as previous generations of cellular networks have adopted nearly the same RAN architecture, limiting the potential for further performance improvement. We have investigated the existing paradigms and formulated the design objectives for the next generation RAN architecture. We have also presented FD-RAN as a promising RAN design that not only incorporates the advantages of existing RANs but also fully decouples the RAN. We have presented surveys on the promising technologies that can be enabled by transforming the underlying RAN architectures, so as to boost the performance of 6G. Case studies on the key technologies in FD-RAN are also provided to demonstrate its core mechanisms. Future directions, especially the integration of FD-RAN and emerging services of 6G are further discussed. We hope that this survey will encourage renovation of 6G RAN architecture, thereby facilitating the implementation of state-of-the-art technologies.

References

[1] N. Apostolakis, M. Gramaglia, and P. Serrano, "Design and validation of an open source cloud native mobile network," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 66–72, Nov. 2022.

[2] R. Schmidt and N. Nikaein, "RAN engine: Service-oriented RAN through Containerized micro-services," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 1, pp. 469–481, Mar. 2021.

[3] "A flagship for 6G vision and intelligent fabric of technology enablers connecting human, physical, and digital worlds." Accessed: May 17, 2023. [Online]. Available: https://hexa-x.eu/

[4] "Building the foundation for north American leadership in 6G and beyond." Accessed: May 17, 2023. [Online]. Available: https://nextgalliance.org/#

[5] "WorldŠs first 6G research programme." Accessed: May 17, 2023. [Online]. Available: https://www.6gflagship.com/

[6] C.-X. Wang et al., "On the road to 6G: Visions, requirements, key technologies, and Testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quart., 2023.

[7] D. C. Nguyen et al., "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, Jan. 2022.

[8] E. Bertin, N. Crespi, and T. Magedanz, *Shaping Future 6G Networks: Needs, Impacts, and Technologies*. Hoboken, NJ, USA: Wiley, 2021.

[9] L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-twin-enabled 6G: Vision, architectural trends, and future directions," *IEEE Commun. Mag.*, vol. 60, no. 1, pp. 74–80, Jan. 2022.

[10] A. H. Khan et al., "Blockchain and 6G: The future of secure and ubiquitous communication," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 194–201, Feb. 2022.

[11] L. Li et al., "mmWave communications for 5G: Implementation challenges and advances," *Sci. China Inf. Sci.*, vol. 61, pp. 1–19, Feb. 2018.

[12] M. Gomez, M. Weiss, and P. Krishnamurthy, "Improving liquidity in secondary spectrum markets: Virtualizing spectrum for fungibility," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 2, pp. 252–266, Jun. 2019.

[13] (Ericsson, Stockholm, Sweden). *Breaking the Energy Curve*. Accessed: Feb. 3, 2023. [Online]. Available: https://www.ericsson.com/495d5c/assets/local/about-ericsson/sustainability-and-corporate-responsibility/documents/2020/breaking-the-energy-curve-report.pdf,

[14] Y. Han, S. Jin, C.-K. Wen, and T. Q. S. Quek, "Localization and channel reconstruction for extra large RIS-assisted massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1011–1025, Aug. 2022.

[15] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, "Non-stationarities in extra-large-scale massive MIMO," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 74–80, Aug. 2020.

[16] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 990–1002, Sep. 2020.

[17] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," 2023, *arXiv:2303.16129*.

[18] Y. Wu, Y. Song, T. Wang, L. Qian, and T. Q. S. Quek, "Non-orthogonal multiple access assisted federated learning via wireless power transfer: A cost-efficient approach," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2853–2869, Apr. 2022.

[19] Y. Yang et al., "6G network AI architecture for everyone-centric customized services," *IEEE Netw.*, vol. 37, no. 5, pp. 71–80, Sep. 2023.

[20] J. Chen, B. Qian, Y. Xu, H. Zhou, and X. Shen, "Towards user-centric resource allocation for 6G: An economic perspective," *IEEE Netw.*, vol. 37, no. 2, pp. 254–261, Mar./Apr. 2023.

[21] C. Sexton, N. J. Kaminski, J. M. Marquez-Barja, N. Marchetti, and L. A. DaSilva, "5G: Adaptable networks enabled by versatile radio access technologies," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 688–720, 2nd Quart., 2017.

[22] L. M. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 146–172, 1st Quart., 2018.

[23] A. Dogra, R. K. Jha, and S. Jain, "A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies," *IEEE Access*, vol. 9, pp. 67512–67547, 2021.

[24] A. Checko et al., "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[25] M. F. Hossain, A. U. Mahin, T. Debnath, F. B. Mosharrof, and K. Z. Islam, "Recent research in cloud radio access network (C-RAN) for 5G cellular systems—A survey," *J. Netw. Comput. Appl.*, vol. 139, pp. 31–48, Aug. 2019.

[26] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 668–695, 2nd Quart., 2021.

[27] B. Agarwal, M. A. Togou, M. Marco, and G.-M. Muntean, "A comprehensive survey on radio resource management in 5G HetNets: Current solutions, future trends and open issues," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2495–2534, 4th Quart., 2022.

[28] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[29] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[30] Y. Shi, E. Alsusa, and M. W. Baidas, "A survey on downlink–uplink decoupled access: Advances, challenges, and open problems," *Comput. Netw.*, vol. 213, Aug. 2022, Art. no. 109040.

[31] F. Z. Morais, C. A. da Costa, A. M. Alberti, C. B. Both, and R. da Rosa Righi, "When SDN meets C-RAN: A survey exploring multi-point coordination, interference, and performance," *J. Netw. Comput. Appl.*, vol. 162, Jul. 2020, Art. no. 102655.

[32] R. T. Rodoshi and W. Choi, "A survey on applications of deep learning in cloud radio access network," *IEEE Access*, vol. 9, pp. 61972–61997, 2021.

[33] N.-N. Dao et al., "Survey on aerial radio access networks: Toward a comprehensive 6G access infrastructure," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1193–1225, 2nd Quart., 2021.

[34] M. Polese, L. Bonati, S. D'oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1376–1411, 2nd Quart., 2023.

[35] Q. Yu et al., "A fully-decoupled RAN architecture for 6G inspired by neurotransmission," *J. Commun. Inf. Netw.*, vol. 4, no. 4, pp. 15–23, Dec. 2019.

[36] E. R. Kandel et al., *Principles of Neural Science*, vol. 4, New York, NY, USA: McGraw-hill, 2000.

[37] (Huawei Manuf. Co., Shenzhen, China). *Ten Wireless Industry Trends for Mobile 2030*. (2021). [Online]. Available: https://www.huawei.com/cn/huaweitech/industry-insights/outlook/mobile-2030-10-wireless-industry-trends

[38] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM J. Res. Develop.*, vol. 54, no. 1, pp. 4:1–4:12, Jan./Feb. 2010.

[39] *C-RAN: The Road Towards Green RAN*, China Mobile Res. Inst., Beijing, China, 2011.
[40] M. Masoudi, S. S. Lisi, and C. Cavdar, "Cost-effective migration toward virtualized C-RAN with scalable fronthaul design," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5100–5110, Dec. 2020.
[41] V. Suryaprakash, P. Rost, and G. Fettweis, "Are heterogeneous cloud-based radio access networks cost effective?" *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2239–2251, Oct. 2015.
[42] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour, "Distributed inter-BS cooperation aided energy efficient load balancing for cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5929–5939, Nov. 2013.
[43] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-RAN architecture for future cellular network," in *Proc. Future Netw. Mobile Summit (FutureNetw)*, 2012, pp. 1–8.
[44] S. Bhaumik et al., "CloudIQ: A framework for processing base stations in a data center," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw.*, 2012, pp. 125–136.
[45] G. Li et al., "Architecture of GPP based, scalable, large-scale C-RAN BBU pool," in *Proc. IEEE Globecom Workshops*, 2012, pp. 267–272.
[46] *ZTE Green Technology Innovations*, ZTE Telecoms Equip. Corp., Shenzhen, China, 2011.
[47] "Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN; Version 12.1.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.932, Mar. 2013.
[48] L. Chen, J. Wu, X. Zhang, and G. Zhou, "TARCO: Two-stage auction for D2D relay aided computation resource allocation in HetNet," *IEEE Trans. Services Comput.*, vol. 14, no. 1, pp. 286–299, Jan./Feb. 2021.
[49] M. Sepulcre and J. Gozalvez, "Heterogeneous V2V communications in multi-link and multi-RAT vehicular networks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 1, pp. 162–173, Jan. 2021.
[50] T. Sylla, L. Mendiboure, S. Maaloul, H. Aniss, M. A. Chalouf, and S. Delbruel, "Multi-connectivity for 5G networks and beyond: A survey," *Sensors*, vol. 22, no. 19, p. 7591, Jan. 2022.
[51] T. Z. Oo, N. H. Tran, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Offloading in HetNet: A coordination of interference mitigation, user association, and resource allocation," *IEEE Trans. Mobile Comput.*, vol. 16, no. 8, pp. 2276–2291, Aug. 2017.
[52] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.
[53] W. Xia, J. Zhang, T. Q. S. Quek, S. Jin, and H. Zhu, "Joint optimization of fronthaul compression and bandwidth allocation in uplink H-CRAN with large system analysis," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6556–6569, Dec. 2018.
[54] E. Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy-efficient multi-stream carrier aggregation for heterogeneous networks in 5G wireless systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7432–7443, Nov. 2016.
[55] H. Cui and F. You, "User-centric resource scheduling for dual-connectivity communications," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3659–3663, Nov. 2021.
[56] J. Chen, Z. Ma, Y. Liu, J. Jia, and X. Wang, "Energy efficient resource allocation for MSCA enabled CoMP in HetNets," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2965–2978, Mar. 2022.
[57] F. Boccardi et al., "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 110–117, Mar. 2016.
[58] J. Li, X. Wang, Z. Li, H. Wang, and L. Li, "Energy efficiency optimization based on eICIC for wireless heterogeneous networks," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10166–10176, Dec. 2019.
[59] A. Alnoman and A. Anpalagan, "Computing-aware base station sleeping mechanism in H-CRAN-cloud-edge networks," *IEEE Trans. Cloud Comput.*, vol. 9, no. 3, pp. 958–967, Jul./Sep. 2021.
[60] "Study on small cell enhancements for E-UTRA and E-UTRAN; higher layer aspects; Version 12.0.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.842, Dec. 2013.
[61] "Evolved universal terrestrial radio access (E-UTRA) and NR; multi-connectivity; stage 2; Version 17.2.0," 3GPP, Sophia Antipolis, France, Rep. TR-37.340, Sep. 2022.
[62] "Evolved universal terrestrial radio access (E-UTRA); LTE advanced inter-band carrier aggregation (CA) Rel-14 for 3 down link (DL) / 1 up link (UL); Version 14.0.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.714, Jun. 2017.
[63] "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2; Version 10.5.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.300, Sep. 2011.
[64] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2015, pp. 201–205.
[65] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–13, 2019. [Online]. Available: https://link.springer.com/article/10.1186/s13638-019-1507-0#citeas
[66] V. Ranjbar, A. Girycki, M. A. Rahman, S. Pollin, M. Moonen, and E. Vinogradov, "Cell-free MIMO support in the O-RAN architecture: A PHY layer perspective for 5G and beyond networks," *IEEE Commun. Stand. Mag.*, vol. 6, no. 1, pp. 28–34, Mar. 2022.
[67] F. Qamar, K. B. Dimyati, M. N. Hindia, K. A. B. Noordin, and A. M. Al-Samman, "A comprehensive review on coordinated multi-point operation for LTE-A," *Comput. Netw.*, vol. 123, pp. 19–37, Aug. 2017.
[68] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 1–30, 1st Quart., 2022.
[69] Q. Yu, J. Ren, H. Zhou, and W. Zhang, "A cybertwin based network architecture for 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.
[70] Q. Yu, J. Ren, Y. Fu, Y. Li, and W. Zhang, "Cybertwin: An origin of next generation network architecture," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 111–117, Dec. 2019.
[71] D. Hu, J. Chen, H. Zhou, K. Yu, B. Qian, and W. Xu, "Leveraging blockchain for multi-operator access sharing management in Internet of Vehicles," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2774–2787, Mar. 2022.
[72] J. Liu, J. Chen, C. He, and H. Zhou, "Leveraging load-aware dynamic pricing for cell-level demand-supply equilibrium," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6902–6906, May 2023.
[73] Y. Sun, K. Yu, Y. Xu, H. Zhou, and X. Shen, "Flexible base station sleeping and resource cooperation enabled green fully-decoupled RAN," 2023, *arXiv:2312.05517*.
[74] J. Zhao et al., "Fully-decoupled radio access networks: A resilient uplink base stations cooperative reception framework," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5096–5110, Aug. 2023.
[75] K. Yu, H. Zhou, Z. Tang, X. Shen, and F. Hou, "Deep reinforcement learning-based RAN slicing for UL/DL decoupled cellular V2X," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3523–3535, May 2022.
[76] L. Jiao, K. Yu, Y. Xu, T. Zhang, H. Zhou, and X. Shen, "Spectral efficiency analysis of uplink-downlink decoupled access in C-V2X networks," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2022, pp. 2062–2067.
[77] Y. Xu, B. Qian, K. Yu, T. Ma, L. Zhao, and H. Zhou, "Federated learning over fully-decoupled RAN architecture for two-tier computing acceleration," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 789–801, Mar. 2023.
[78] T. Liu, H. Zhou, J. Li, F. Shu, and Z. Han, "Uplink and downlink decoupled 5G/B5G vehicular networks: A federated learning assisted client selection method," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2280–2292, Feb. 2023.
[79] M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-connectivity as an enabler for reliable low latency communications—An overview," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 156–169, 1st Quart., 2020.
[80] A. Alorainy and M. J. Hossain, "Cross-layer performance analysis of downlink multi-flow carrier aggregation in heterogeneous networks," *IEEE Access*, vol. 7, pp. 23303–23318, 2019.
[81] S. Kim, "Two-level game based spectrum allocation scheme for multi-flow carrier aggregation technique," *IEEE Access*, vol. 8, pp. 89291–89299, 2020.
[82] F. Foukalas, R. Shakeri, and T. Khattab, "Distributed power allocation for multi-flow carrier aggregation in heterogeneous cognitive cellular networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2486–2498, Apr. 2018.
[83] B. Soret, H. Wang, K. I. Pedersen, and C. Rosa, "Multicell cooperation for LTE-advanced heterogeneous network scenarios," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 27–34, Feb. 2013.

[84] C. Rosa et al., "Dual connectivity for LTE small cell evolution: Functionality and performance aspects," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 137–143, Jun. 2016.

[85] J. Xue, K. Yu, T. Zhang, H. Zhou, L. Zhao, and X. Shen, "Cooperative deep reinforcement learning enabled power allocation for packet duplication URLLC in multi-connectivity vehicular networks," *IEEE Trans. Mobile Comput.*, early access, Jan. 3, 2024, doi: 10.1109/TMC.2023.3347580.

[86] M. G. Kibria, K. Nguyen, G. P. Villardi, K. Ishizu, and F. Kojima, "Next generation new radio small cell enhancement: Architectural options, functionality and performance aspects," *IEEE Wireless Commun.*, vol. 25, no. 4, pp. 120–128, Aug. 2018.

[87] O. N. C. Yilmaz, O. Teyeb, and A. Orsino, "Overview of LTE-NR dual connectivity," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 138–144, Jun. 2019.

[88] M. Agiwal, H. Kwon, S. Park, and H. Jin, "A survey on 4G-5G dual connectivity: Road to 5G implementation," *IEEE Access*, vol. 9, pp. 16193–16210, 2021.

[89] G. Liu, Y. Huang, Z. Chen, L. Liu, Q. Wang, and N. Li, "5G deployment: Standalone vs. non-standalone from the operator perspective," *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 83–89, Nov. 2020.

[90] M. Shafi et al., "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[91] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 743–764, 2nd Quart., 2017.

[92] D. Lee et al., "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[93] "Evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects; Version 11.2.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.814, Mar. 2010.

[94] "Requirements for further advancements for evolved universal terrestrial radio access (E-UTRA) (LTE-advanced); Version 9.0.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.913, Dec. 2009.

[95] P. Marsch and G. P. Fettweis, *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[96] "Coordinated multi-point operation for LTE physical layer aspects; Version 11.2.0," 3GPP, Sophia Antipolis, France, Rep. TR-36.819, Dec. 2013.

[97] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Optimal dynamic point selection for power minimization in multiuser downlink CoMP," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 619–633, Jan. 2017.

[98] Y. Al-Eryani, E. Hossain, and D. I. Kim, "Generalized coordinated multipoint (GCoMP)-enabled NOMA: Outage, capacity, and power allocation," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7923–7936, Nov. 2019.

[99] Q. Cui et al., "Evolution of limited-feedback CoMP systems from 4G to 5G: CoMP features and limited-feedback approaches," *IEEE Veh. Technol. Mag.*, vol. 9, no. 3, pp. 94–103, Sep. 2014.

[100] S. Gulia, A. Ahmad, S. Singh, and M. D. Gupta, "Interference management in backhaul constrained 5G HetNets through coordinated multipoint," *Comput. Elect. Eng.*, vol. 100, May 2022, Art. no. 107982.

[101] D. Marabissi, G. Bartoli, R. Fantacci, and M. Pucci, "An Optimized CoMP transmission for a heterogeneous network using eICIC approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8230–8239, Oct. 2016.

[102] Z. Zhang, G. Yang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Heterogeneous ultradense networks with NOMA: System architecture, coordination framework, and performance evaluation," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 110–120, Jun. 2018.

[103] B. Li, Y. Dai, Z. Dong, E. Panayirci, H. Jiang, and H. Jiang, "Energy-efficient resources allocation with Millimeter-wave massive MIMO in ultra dense HetNets by SWIPT and CoMP," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4435–4451, Jul. 2021.

[104] J. Tang, A. Shojaeifard, D. K. C. So, K.-K. Wong, and N. Zhao, "Energy efficiency optimization for CoMP-SWIPT heterogeneous networks," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6368–6383, Dec. 2018.

[105] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, Sep. 2018.

[106] S. Rezvani, N. M. Yamchi, M. R. Javan, and E. A. Jorswieck, "Resource allocation in virtualized CoMP-NOMA HetNets: Multi-connectivity for joint transmission," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4172–4185, Jun. 2021.

[107] "IMT traffic estimates for the years 2020 to 2030," ITU-Rec. M.2370, Int. Telecommun. Union, Geneva, Switzerland, 2015.

[108] J. Oueis and E. C. Strinati, "Uplink traffic in future mobile networks: Pulling the alarm," in *Proc. 11th Int. Conf. Cogn. Radio Orient. Wireless Netw. (CROWNCOM)*, 2016, pp. 583–593.

[109] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62367–62414, 2020.

[110] A. Ullah, Z. H. Abbas, F. Muhammad, G. Abbas, and S. Kim, "Uplink performance analysis of user-centric small cell aided dense HCNets with uplink-downlink decoupling," *IEEE Access*, vol. 8, pp. 148460–148474, 2020.

[111] Y. Dhungana and C. Tellambura, "Multichannel analysis of cell range expansion and resource partitioning in two-tier heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2394–2406, Mar. 2016.

[112] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[113] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 154–160, Jul. 2013.

[114] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc. IEEE Glob. Commun. Conf.*, 2014, pp. 1798–1803.

[115] L. Zhang, W. Nie, G. Feng, F.-C. Zheng, and S. Qin, "Uplink performance improvement by decoupling uplink/downlink access in HetNets," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6862–6876, Aug. 2017.

[116] R. Li, K. Luo, T. Jiang, and S. Jin, "Uplink spectral efficiency analysis of decoupled access in multiuser MIMO HetNets," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4289–4302, May 2018.

[117] Z. Sattar, J. V. C. Evangelista, G. Kaddoum, and N. Batani, "Spectral efficiency analysis of the decoupled access for downlink and uplink in two-tier network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4871–4883, May 2019.

[118] Y. Ramamoorthi and A. Kumar, "Energy efficiency in millimeter wave based cellular networks with DUDe and dynamic TDD," in *Proc. Int. Conf. COMmun. Syst. NETw. (COMSNETS)*, 2020, pp. 670–673.

[119] K. Smiljkovikj, L. Gavrilovska, and P. Popovski, "Efficiency analysis of downlink and uplink decoupling in heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, 2015, pp. 125–130.

[120] K. Sun, J. Wu, W. Huang, H. Zhang, H.-Y. Hsieh, and V. C. M. Leung, "Uplink performance improvement for downlink-uplink decoupled HetNets with non-uniform user distribution," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7518–7530, Jul. 2020.

[121] C. Dai, K. Zhu, C. Yi, and E. Hossain, "Decoupled uplink-downlink association in full-duplex cellular networks: A contract-theory approach," *IEEE Trans. Mobile Comput.*, vol. 21, no. 3, pp. 911–925, Mar. 2022.

[122] S. Singh, X. Zhang, and J. G. Andrews, "Joint rate and SINR coverage analysis for decoupled uplink-downlink biased cell associations in HetNets," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5360–5373, Oct. 2015.

[123] M. Bacha, Y. Wu, and B. Clerckx, "Downlink and uplink decoupling in two-tier heterogeneous networks with multi-antenna base stations," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2760–2775, May 2017.

[124] B. Soret, P. Popovski, and K. Stern, "A queueing approach to the latency of decoupled UL/DL with flexible TDD and asymmetric services," *IEEE Wireless Communicatins Lett.*, vol. 8, no. 6, pp. 1704–1708, Dec. 2019.

[125] A. Al-Shuwaili and A. Lawey, "Latency reduction for mobile edge computing in HetNets by uplink and downlink decoupled access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2205–2209, Oct. 2021.

[126] K. Ahmadi, M. Ghassemian, M. Condoluci, and M. Dohler, "X2-based signalling mechanisms for downlink uplink decoupling in next generation communication systems," *IEEE Access*, vol. 10, pp. 88941–88955, 2022.

[127] X. Liu, R. Li, K. Luo, and T. Jiang, "Downlink and uplink decoupling in heterogeneous networks for 5G and beyond," *J. Commun. Inf. Netw.*, vol. 3, no. 2, pp. 1–13, Jun. 2018.

[128] G. Del Galdo and M. Haardt, "Comparison of zero-forcing methods for downlink spatial multiplexing in realistic multi-user MIMO channels," in *Proc. IEEE 59th Veh. Technol. Conf.*, 2004, pp. 299–303.
[129] H. Sung, S.-R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3489–3499, Nov. 2009.
[130] J. C. De Luna Ducoing, Y. Qin, Y. Xue, and K. Nikitopoulos, "Gyre precoding for MU-MIMO systems," *IEEE Commun. Lett.*, vol. 25, no. 8, pp. 2723–2727, Aug. 2021.
[131] B. C. Pandey, S. K. Mohammed, P. Raviteja, Y. Hong, and E. Viterbo, "Low complexity precoding and detection in multi-user massive MIMO OTFS downlink," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4389–4405, May 2021.
[132] V. A. Vaishampayan, "Precoder design for communication-efficient distributed MIMO receivers with controlled peak-average power ratio," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4704–4716, Jul. 2021.
[133] M. K. Arti, "A novel downlink interference alignment method for multi-user MIMO system with no CSIT: A space-time coding approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10937–10948, Oct. 2020.
[134] A. Amiri, S. Rezaie, C. N. Manchon, and E. de Carvalho, "Distributed receiver processing for extra-large MIMO arrays: A message passing approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2654–2667, Apr. 2022.
[135] A. Fengler, O. Musa, P. Jung, and G. Caire, "Pilot-based unsourced random access with a massive MIMO receiver, interference cancellation, and power control," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1522–1534, May 2022.
[136] M. Darabi, A. C. Cirik, and L. Lampe, "Transceiver design in millimeter wave full-duplex multi-user massive MIMO communication systems," *IEEE Access*, vol. 9, pp. 165394–165408, 2021.
[137] J. Du, M. Han, L. Jin, Y. Hua, and X. Li, "Semi-blind receivers for multi-user massive MIMO relay systems based on block Tucker2-PARAFAC tensor model," *IEEE Access*, vol. 8, pp. 32170–32186, 2020.
[138] G. Fodor, S. Fodor, and M. Telek, "MU-MIMO receiver design and performance analysis in time-varying rayleigh fading," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1214–1228, Feb. 2022.
[139] B. P. Maza, G. Dahman, G. Kaddoum, and F. Gagnon, "Average vector-symbol error rate closed-form expression for ML group detection receivers in large MU-MIMO channels with transmit correlation," *IEEE Access*, vol. 8, pp. 45653–45663, 2020.
[140] D. F. Carrera, D. Zabala-Blanco, C. Vargas-Rosales, and C. A. Azurdia-Meza, "Extreme learning machine-based receiver for multi-user massive MIMO systems," *IEEE Commun. Lett.*, vol. 25, no. 2, pp. 484–488, Feb. 2021.
[141] M. Goutay, F. A. Aoudia, J. Hoydis, and J.-M. Gorce, "Machine learning for MU-MIMO receive processing in OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2318–2332, Aug. 2021.
[142] R. Gui, N. M. Balasubramanya, and L. Lampe, "Outage performance analysis of widely linear receivers in uplink multi-user MIMO systems," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6500–6515, Oct. 2021.
[143] H. J. Park and J. W. Lee, "LDPC coded multi-user massive MIMO systems with low-complexity detection," *IEEE Access*, vol. 10, pp. 25296–25308, 2022.
[144] D. Wang, Y. Jiang, J. Hua, X. Gao, and X. You, "Low complexity soft decision equalization for block transmission systems," in *IEEE Int. Conf. Commun.*, vol. 4, 2005, pp. 2372–2376.
[145] M. M. Rahman, C. Despins, and S. Affes, "Analysis of CAPEX and OPEX benefits of wireless access virtualization," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, 2013, pp. 436–440.
[146] N. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," *Comput. Netw.*, vol. 54, no. 5, pp. 862–876, 2010.
[147] X. Wang, P. Krishnamurthy, and D. Tipper, "Wireless network virtualization," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, 2013, pp. 818–822.
[148] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2015.
[149] W. Al-Zubaedi and H. S. Al-Raweshidy, "A parameterized and optimized BBU pool virtualization power model for C-RAN architecture," in *Proc. Int. Conf. Smart Technol.*, 2017, pp. 38–43.
[150] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 708–769, 1st Quart., 2018.
[151] C. Liang and F. R. Yu, "Wireless virtualization for next generation mobile cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 61–69, Feb. 2015.
[152] S. Das, F. Slyne, and M. Ruffini, "Optimal slicing of virtualized passive optical networks to support dense deployment of cloud-RAN and multi-access edge computing," *IEEE Netw.*, vol. 36, no. 2, pp. 131–138, Mar./Apr. 2022.
[153] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Comput. Netw.*, vol. 106, pp. 17–48, Sep. 2016.
[154] G. C. Valastro, D. Panno, and S. Riolo, "A SDN/NFV based C-RAN architecture for 5G mobile networks," in *Proc. Int. Conf. Select. Topics Mobile Wireless Netw. (MoWNeT)*, 2018, pp. 1–8.
[155] E. Ahvar, S. Ahvar, S. M. Raza, J. M. S. Vilchez, and G. M. Lee, "Next generation of SDN in cloud-fog for 5G and beyond-enabled applications: Opportunities and challenges," *Network*, vol. 1, no. 1, pp. 28–49, 2021.
[156] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar./Apr. 2015.
[157] A. G. Dalla-Costa, L. Bondan, J. A. Wickboldt, C. B. Both, and L. Z. Granville, "Orchestra: A Customizable split-aware NFV orchestrator for dynamic cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 6, pp. 1014–1024, Jun. 2020.
[158] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.
[159] H. Wen, P. K. Tiwary, and T. Le-Ngoc, "Current trends and perspectives in wireless virtualization," in *Proc. Int. Conf. Select. Topics Mobile Wireless Netw. (MoWNeT)*, 2013, pp. 62–67.
[160] W. Ejaz, S. K. Sharma, S. Saadat, M. Naeem, A. Anpalagan, and N. A. Chughtai, "A comprehensive survey on resource allocation for CRAN in 5G and beyond networks," *J. Netw. Comput. Appl.*, vol. 160, Jun. 2020, Art. no. 102638.
[161] W. Xia, T. Q. S. Quek, J. Zhang, S. Jin, and H. Zhu, "Programmable hierarchical C-RAN: From task scheduling to resource allocation," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 2003–2016, Mar. 2019.
[162] Z. Becvar, P. Mach, M. Elfiky, and M. Sakamoto, "Hierarchical scheduling for suppression of fronthaul delay in C-RAN with dynamic functional split," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 95–101, Apr. 2021.
[163] A. A. Samarneh and A. Y. Alma'aitah, "A scheduling algorithm for adaptive C-RAN architecture," in *Proc. 13th Int. Conf. Inf. Commun. Syst. (ICICS)*, 2022, pp. 82–86.
[164] N. Budhdev, A. Maity, M. C. Chan, and T. Mitra, "Load balancing for a user-level virtualized 5G cloud-RAN," in *Proc. 17th ACM Workshop Mobil. Evol. Internet Archit.*, 2022, pp. 1–6.
[165] M. Mouawad, F. Mah, and Z. Dziong, "RRH-sector selection and load balancing based on MDP and dynamic RRH-sector-BBU mapping in C-RAN," *Comput. Netw.*, vol. 215, Oct. 2022, Art. no. 109192.
[166] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint precoding and RRH selection for user-centric green MIMO C-RAN," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2891–2906, May 2017.
[167] S. Ali, A. Haider, M. Rahman, M. Sohail, and Y. B. Zikria, "Deep learning (DL) based joint resource allocation and RRH association in 5G-multi-tier networks," *IEEE Access*, vol. 9, pp. 118357–118366, 2021.
[168] N. Kumar and A. Ahmad, "Cooperative evolution of support vector machine empowered knowledge-based radio resource management for 5G C-RAN," *Ad Hoc Netw.*, vol. 136, Nov. 2022, Art. no. 102960.
[169] J. Khan and L. Jacob, "Resource allocation for CoMP enabled URLLC in 5G C-RAN architecture," *IEEE Syst. J.*, vol. 15, no. 4, pp. 4864–4875, Dec. 2021.
[170] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.
[171] M. Setayesh, S. Bahrami, and V. W. Wong, "Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.
[172] A. Younis, T. X. Tran, and D. Pompili, "Energy-efficient resource allocation in C-RANs with capacity-limited fronthaul," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 473–487, Feb. 2021.

[173] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1857–1868, Jun. 2018.

[174] A. F. Zavleh and H. Bakhshi, "Downlink resource allocation to total system transmit power minimization in SCMA-based systems for cloud-RAN in 5G networks," *Telecommun. Syst.*, vol. 81, no. 4, pp. 575–590, 2022.

[175] K. Wang, K. Yang, and C. S. Magurawalage, "Joint energy minimization and resource allocation in C-RAN with mobile cloud," *IEEE Trans. Cloud Comput.*, vol. 6, no. 3, pp. 760–770, Jul.–Sep. 2018.

[176] J. Chen, J. Gong, X. Chen, and T. Hsu, "Resource joint allocation scheme based on network slicing under C-RAN architecture," in *Proc. Int. Conf. Internet Things Service*, 2020, pp. 23–35.

[177] Y.-S. Chen, C.-S. Hsu, and H.-C. Hung, "Optimizing communication and computational resource allocations in network slicing using twin-GAN-based DRL for 5G hybrid C-RAN," *Comput. Commun.*, vol. 200, pp. 66–85, Feb. 2023.

[178] A. Mohajer, F. Sorouri, A. Mirzaei, A. Ziaeddini, K. J. Rad, and M. Bavaghar, "Energy-aware hierarchical resource management and backhaul traffic optimization in heterogeneous cellular networks," *IEEE Syst. J.*, vol. 16, no. 4, pp. 5188–5199, Dec. 2022.

[179] Y. Xu, M. Yang, Y. Yang, Y. Ye, R. Q. Hu, and D. Li, "Max-min energy-efficient optimization for cognitive heterogeneous networks with spectrum sensing errors and channel uncertainties," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1113–1117, Jun. 2022.

[180] X. Liu, H. Zhang, K. Long, A. Nallanathan, and V. C. M. Leung, "Energy efficient user association, resource allocation and caching deployment in fog radio access networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1846–1856, Feb. 2022.

[181] U. Ghafoor, H. Z. Khan, M. Ali, A. M. Siddiqui, M. Naeem, and I. Rashid, "Energy efficient resource allocation for H-NOMA assisted B5G HetNets," *IEEE Access*, vol. 10, pp. 91699–91711, 2022.

[182] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. K. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12244–12258, Dec. 2018.

[183] P. Qin, Y. Fu, X. Zhao, K. Wu, J. Liu, and M. Wang, "Optimal task offloading and resource allocation for C-NOMA heterogeneous air-ground integrated power Internet of Things networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9276–9292, Nov. 2022.

[184] H. Xiao, W. Zhang, and A. T. Chronopoulos, "Joint subchannel and power allocation for energy efficiency optimization in NOMA heterogeneous networks with energy harvesting," *IEEE Syst. J.*, vol. 16, no. 3, pp. 4904–4915, Sep. 2022.

[185] M. Moghimi, A. Zakeri, M. R. Javan, N. Mokari, and D. W. K. Ng, "Joint radio resource allocation and cooperative caching in PD-NOMA-based HetNets," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2029–2044, Jun. 2022.

[186] Y. Zhang, H. Zhang, H. Zhou, K. Long, and G. K. Karagiannidis, "Resource allocation in terrestrial-satellite-based next generation multiple access networks with interference cooperation," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1210–1221, Apr. 2022.

[187] W. Jiang, K. Huang, Y. Chen, X. Sun, J. Yang, and K. Zhao, "A joint design for multi-band heterogeneous networks when deploying reconfigurable intelligent surface," *J. Commun. Netw.*, vol. 24, no. 5, pp. 613–623, Oct. 2022.

[188] K. Khawam, S. Lahoud, M. E. Helou, S. Martin, and F. Gang, "Coordinated framework for spectrum allocation and user association in 5G HetNets with mmWave," *IEEE Trans. Mobile Comput.*, vol. 21, no. 4, pp. 1226–1243, Apr. 2022.

[189] J. Du, C. Jiang, A. Benslimane, S. Guo, and Y. Ren, "SDN-based resource allocation in edge and cloud computing systems: An evolutionary Stackelberg differential game approach," *IEEE/ACM Trans. Netw.*, vol. 30, no. 4, pp. 1613–1628, Aug. 2022.

[190] S. Ghosh, D. De, and P. Deb, "E2Beam: Energy efficient beam allocation in 5G HetNet using cooperative game," in *Proc. IEEE Int. Women Eng. (WIE) Conf. Electr. Comput. Eng. (WIECON-ECE)*, 2020, pp. 219–222.

[191] M. Amine, A. Kobbane, and J. Ben-Othman, "New network slicing scheme for UE association solution in 5G ultra dense HetNets," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[192] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2773–2782, Dec. 2020.

[193] H. Yang, J. Zhao, K.-Y. Lam, Z. Xiong, Q. Wu, and L. Xiao, "Distributed deep reinforcement learning-based spectrum and power allocation for heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6935–6948, Sep. 2022.

[194] A. Alwarafy, B. S. Çiftler, M. Abdallah, M. Hamdi, and N. Al-Dhahir, "Hierarchical multi-agent DRL-based framework for joint multi-RAT assignment and dynamic resource allocation in next-generation HetNets," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2481–2494, Jul./Aug. 2022.

[195] J. Jang and H. J. Yang, "Recurrent neural network-based user association and power control in dynamic HetNets," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9674–9689, Sep. 2022.

[196] M. Sana, A. De Domenico, and E. Calvanese Strinati, "Multi-agent deep reinforcement learning based user association for dense mmWave networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. Waikoloa, HI, USA, 2019, pp. 1–6.

[197] J. Jang and H. J. Yang, "Deep reinforcement learning-based resource allocation and power control in small cells with limited information exchange," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13768–13783, Nov. 2020.

[198] M. U. A. Siddiqui, F. Qamar, F. Ahmed, Q. N. Nguyen, and R. Hassan, "Interference management in 5G and beyond network: requirements, challenges and future directions," *IEEE Access*, vol. 9, pp. 68932–68965, 2021.

[199] M. A. Adedoyin and O. E. Falowo, "Combination of ultra-dense networks and other 5G enabling technologies: A survey," *IEEE Access*, vol. 8, pp. 22893–22932, 2020.

[200] M. Waqas et al., "A comprehensive survey on mobility-aware D2D communications: Principles, practice and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1863–1886, 3rd Quart., 2020.

[201] F. Liu, H. Zhao, and Y. Tang, "An eigen domain interference rejection combining algorithm for narrowband interference suppression," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 813–816, May 2014.

[202] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Q. Quek, and J. Zhang, "Enhanced Intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, Jun. 2011.

[203] L. Suo, H. Li, S. Zhang, and J. Li, "Successive interference cancellation and alignment in K-user MIMO interference channels with partial unidirectional strong interference," *China Commun.*, vol. 19, no. 2, pp. 118–130, Feb. 2022.

[204] S.-J. Kim, I. Ban, J. Park, and Y. Na, "SINR of machine-type communication in HetNets: Interference avoidance and CoMP schemes," in *Proc. TRON Symp. (TRONSHOW)*, 2022, pp. 1–5.

[205] H. Zhang, K. Niu, J. Xu, J. Dai, and J. Zhang, "Iterative SIC-based multiuser detection for uplink heterogeneous NOMA system," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 94–99.

[206] S. Rezvani, E. A. Jorswieck, N. M. Yamchi, and M. R. Javan, "Optimal SIC ordering and power allocation in downlink multi-cell NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3553–3569, Jun. 2022.

[207] M. Venkatesan, A. Kulkarni, R. Menon, and S. Prasad, "Interference mitigation approach using massive MIMO towards 5G networks," in *Proc. 2nd Asian Conf. Innov. Technol. (ASIANCON)*, 2022, pp. 1–5.

[208] C. He, G. Y. Li, F.-C. Zheng, and X. You, "Power allocation criteria for distributed antenna systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5083–5090, Nov. 2015.

[209] A. Mario, B. Stefano, Z. Alessio, and D. Ciro, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 651–663, Sep. 2019.

[210] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao "Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 971–987, Dec. 2019.

[211] M. Alageli, A. Ikhlef, F. Alsifiany, M. A. M. Abdullah, G. Chen, and J. Chambers, "Optimal downlink transmission for cell-free SWIPT massive MIMO systems with active eavesdropping," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1983–1998, 2020.

[212] H. V. Nguyen et al., "On the spectral and energy efficiencies of full-duplex cell-free massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1698–1718, Aug. 2020.

[213] T. K. Nguyen, H. H. Nguyen, and H. D. Tuan "Max-min QoS power control in generalized cell-free massive MIMO-NOMA with optimal backhaul combining," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10949–10964, Oct. 2020.

[214] Q. Jiahua, X. Kui, X. Xiaochen, S. Zhexian, and X. Wei, "Downlink power optimization for cell-free massive MIMO over spatially correlated Rayleigh fading channels," *IEEE Access*, vol. 8, pp. 56214–56227, 2020.

[215] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6798–6812, Oct. 2020.

[216] W. Xinhua, A. Alexei, and W. Xiaodong, "Wirelessly powered cell-free IoT: Analysis and optimization," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8384–8396, Sep. 2020.

[217] Y. Zhao, I. G. Niemegeers, and S. H. De Groot, "Power allocation in cell-free massive MIMO: A deep learning method," *IEEE Access*, vol. 8, pp. 87185–87200, 2020.

[218] B. Manijeh et al., "Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 223–245, Jan. 2021.

[219] D. O. Tugfe and B. Emil, "Joint power control and LSFD for wireless-powered cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1756–1769, Mar. 2021.

[220] S.-N. Jin, D.-W. Yue, and H. H. Nguyen, "Spectral and energy efficiency in cell-free massive MIMO systems over correlated Rician fading," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2822–2833, Jun. 2021.

[221] F. Tan, P. Wu, Y.-C. Wu, and M. Xia, "Energy-efficient non-orthogonal multicast and unicast transmission of cell-free massive MIMO systems with SWIPT," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 949–968, Apr. 2021.

[222] Y. Zhang, Y. Cheng, M. Zhou, L. Yang, and H. Zhu, "Analysis of uplink cell-free massive MIMO system with mixed-ADC/DAC receiver," *IEEE Syst. J.*, vol. 15, no. 4, pp. 5162–5173, Dec. 2021.

[223] H. Yun et al., "Energy efficient power allocation for cell-free mmWave massive MIMO with hybrid precoder," *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 394–398, Feb. 2022.

[224] T. C. Mai, H. Q. Ngo, and L.-N. Tran "Energy efficiency maximization in large-scale cell-free massive MIMO: A projected gradient approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6357–6371, Aug. 2022.

[225] M. Makhanbet, T. Lv, W. Ni, and M. Orynbet, "Energy-delay-aware power control for reliable transmission of dynamic cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 276–290, Jan. 2022.

[226] V. Tentu, E. Sharma, D. N. Amudala, and R. Budhiraja "UAV-enabled hardware-impaired spatially correlated cell-free massive MIMO systems: Analysis and energy efficiency optimization," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2722–2741, Apr. 2022.

[227] H. D. Tuan, A. A. Nasir, H. Q. Ngo, E. Dutkiewicz, and H. V. Poor, "Scalable user rate and energy-efficiency optimization in cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6050–6065, Sep. 2022.

[228] J. Zhang, J. Fan, J. Zhang, D. W. K. Ng, Q. Sun, and B. Ai, "Performance analysis and optimization of NOMA-based cell-free massive MIMO for IoT," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9625–9639, Jun. 2022.

[229] Y. Zhang, W. Xia, H. Zhao, W. Xu, K.-K. Wong, and L. Yang, "Cell-free IoT networks with SWIPT: Performance analysis and power control," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13780–13793, Aug. 2022.

[230] K. Yu et al., "Fully-decoupled radio access networks: A flexible downlink multi-connectivity and dynamic resource cooperation framework," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 4202–4214, Jun. 2023.

[231] B. Qian et al., "Enabling fully-decoupled radio access with elastic resource allocation," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 4, pp. 1025–1040, Aug. 2023.

[232] *Study on 3D Channel Model for LTE, Version 12.7.0*, 3GPP Standard TS 36.873, 2017.

[233] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[234] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.

[235] Z. Wei et al., "Integrated sensing and communication signals towards 5G-A and 6G: A survey," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11068–11092, Jul. 2023.

[236] W. Zhou, R. Zhang, G. Chen, and W. Wu, "Integrated sensing and communication waveform design: A survey," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1930–1949, 2022.

[237] A. Liu et al., "A survey on fundamental limits of integrated sensing and communication," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 994–1034, 2nd Quart., 2022.

[238] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[239] H. Cui et al., "Space-air-ground integrated network (SAGIN) for 6G: Requirements, architecture and challenges," *China Commun.*, vol. 19, no. 2, pp. 90–108, Feb. 2022.

[240] X. Qin, T. Ma, Z. Tang, X. Zhang, H. Zhou, and L. Zhao, "Service-aware resource orchestration in ultra-dense LEO satellite-terrestrial integrated 6G: A service function chain approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6003–6017, Sep. 2023.

[241] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[242] "Stable diffusion 3." Accessed: Jun. 20, 2023. [Online]. Available: https://stability.ai/stablediffusion/

[243] Z. Liu et al., "Leveraging deep reinforcement learning for geolocation-based MIMO transmission in FD-RAN," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2023, pp. 1–6.

[244] M. Hosseinian, J. P. Choi, S.-H. Chang, and J. Lee, "Review of 5G NTN standards development and technical challenges for satellite integration with the 5G network," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 8, pp. 22–31, Aug. 2021.

[245] H. Yang, Z. Xiong, J. Zhao, D. Niyato, C. Yuen, and R. Deng, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2977–2990, May 2021.

[246] H. Zhou, Y. Deng, L. Feltrin, and A. Höglund, "Analyzing novel grant-based and grant-free access schemes for small data transmission," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2805–2819, Apr. 2022.

[247] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 741–755, Mar. 2021.

[248] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 629–646, 1st Quart., 2018.

[249] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[250] "Study on downlink multiuser superposition transmission (MUST) for LTE, Version 13.0.0," 3GPP, Sophia Antipolis, France, Rep. 36.859, Jan. 2016.

[251] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "TUBE: Time-dependent pricing for mobile data," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protoc. Comput. Commun.*, 2012, pp. 247–258.

[252] N. C. Luong, P. Wang, D. Niyato, Y.-C. Liang, Z. Han, and F. Hou, "Applications of economic and pricing models for resource management in 5G wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3298–3339, 4th Quart., 2018.

Jiacheng Chen (Member, IEEE) received the Ph.D. degree in information and communications engineering from Shanghai Jiao Tong University, Shanghai, China, in 2018. From 2015 to 2016, he was a Visiting Scholar with the BBCR Group, University of Waterloo, Canada. He is currently an Assistant Researcher with Peng Cheng Laboratory, Shenzhen, China. His research interests include future network design, AI-enabled 6G network, and resource management. He has won the JCIN Best Paper Award in 2016, the Chinese Institute of Electronics Outstanding Scientific Paper in the Field of Electronic Information in 2020, and the IEEE PIMRC'23 Best Paper Award. He has served as a Guest Editor for IEEE INTERNET OF THINGS JOURNAL and *Journal of Communications and Information Networks*, and the Workshop Co-Chair for IEEE/CIC ICCC from 2021 to 2023.

Xiaohu Liang (Member, IEEE) received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2011, and the Ph.D. degree in information and communication engineering from the PLA University of Science and Technology, Nanjing, China, in 2016. He was invited to visit Lund University for researching the technology of Faster-than-Nyquist signaling. He is a Postdoctoral Research Fellow with the National Mobile Communications Research Laboratory, Southeast University, Nanjing. He is currently with the Faculty in the School of Communication Engineering, Army Engineering University, Nanjing. He has hosted three projects supported by the National Natural Science Foundation of China, the Natural Science Foundation of Jiangsu Province, and China Postdoctoral Science Foundation, respectively. His research interests include non-orthogonal transmission technology, synchronization technology, and Massive MIMO.

Jianzhe Xue (Student Member, IEEE) received the B.S. degree in communication engineering from Xidian University, Xi'an, China in 2021. He is currently pursuing the Ph.D. degree with the School of Electronic Science and Engineering, Nanjing University, China. His current research interests include Internet of Vehicles, orthogonal time frequency space modulation, and machine learning for wireless communications.

Yu Sun (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China in 2022. He is currently pursuing the Ph.D. degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His current research interests include machine learning and optimization for data center, and resource allocation for wireless communications.

Haibo Zhou (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. From 2014 to 2017, he was a Postdoctoral Fellow with the Broadband Communications Research Group, Department of Electrical and Computer Engineering, University of Waterloo. He is currently a Full Professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include resource management and protocol design in B5G/6G networks, vehicular ad hoc networks, and space-air-ground integrated networks. He was a recipient of the 2019 IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award and the 2023 IEEE ComSoc WTC Young Researcher Award. He served as the Track/Symposium Co-Chair for IEEE/CIC ICCC 2019, IEEE VTC-Fall 2020, IEEE VTC-Fall 2021, WCSP 2022, IEEE GLOBECOM 2022, IEEE/CIC ICCC 2024, IEEE ICC 2024, and IEEE GLOBECOM 2024. He is currently an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, *IEEE Network Magazine*, and *Journal of Communications and Information Networks*. He was an IEEE ComSoc Distinguished Lecturer from 2023 to 2024, and an IEEE VTS Distinguished Lecturer 2023 to 2025.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, AI for networks, and vehicular networks. He received the "West Lake Friendship Award" from Zhejiang Province in 2023, the President's Excellence in Research from the University of Waterloo in 2022, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory in 2021, the R.A. Fessenden Award in 2019 from IEEE, Canada, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, the Joseph LoCicero Award in 2015 and the Education Award in 2017 from the IEEE Communications Society, and the Technical Recognition Award from Wireless Communications Technical Committee in 2019, and AHSN Technical Committee in 2013. He has also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier's Research Excellence Award in 2003 from the Province of Ontario, Canada. He is a registered Professional Engineer of Ontario, Canada. He is the President of the IEEE Communications Society. He was the vice president for technical and educational activities, the vice president for publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of IEEE Fellow Selection Committee of the ComSoc. He served as the Editor-in-Chief for the IEEE INTERNET OF THINGS JOURNAL, IEEE NETWORK, and *IET Communications*. He is an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.