IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 3, 2019, accepted May 18, 2019, date of publication May 28, 2019, date of current version June 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919657

A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System

MOHAMMAD ASIF HABIBI[1], MEYSAM NASIMI[1], BIN HAN[1], (Member, IEEE), AND HANS D. SCHOTTEN[1,2], (Member, IEEE)

[1]Institute of Wireless Communication (WiCon), Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany
[2]Intelligent Networks Department, German Research Center for Artificial Intelligence (DFKI GmbH), 67663 Kaiserslautern, Germany

Corresponding author: Mohammad Asif Habibi (asif@eit.uni-kl.de)

This work was supported by the European Union Horizon-2020 Projects 5G-AuRA and 5G-MoNArch under Grant 675806 and Grant 761445.

ABSTRACT The fifth generation (5G) of mobile communication system aims to deliver a ubiquitous mobile service with enhanced quality of service (QoS). It is also expected to enable new use-cases for various vertical industrial applications—such as automobiles, public transportation, medical care, energy, public safety, agriculture, entertainment, manufacturing, and so on. Rapid increases are predicted to occur in user density, traffic volume, and data rate. This calls for novel solutions to the requirements of both mobile users and vertical industries in the next decade. Among various available options, one that appears attractive is to redesign the network architecture—more specifically, to reconstruct the radio access network (RAN). In this paper, we present an inclusive and comprehensive survey on various RAN architectures toward 5G, namely cloud-RAN, heterogeneous cloud-RAN, virtualized cloud-RAN, and fog-RAN. We compare them from various perspectives, such as energy consumption, operations expenditure, resource allocation, spectrum efficiency, system architecture, and network performance. Moreover, we review the key enabling technologies for 5G systems, such as multi-access edge computing, network function virtualization, software-defined networking, and network slicing; and some crucial radio access technologies (RATs), such as millimeter wave, massive multi-input multi-output, device-to-device communication, and massive machine-type communication. Last but not least, we discuss the major research challenges in 5G RAN and 5G RATs and identify several possible directions of future research.

INDEX TERMS 5G, radio access network, network architecture, cloud-RAN, distributed-RAN, fog-RAN, heterogeneous-CRAN, RATs, virtualized-CRAN.

I. INTRODUCTION

The mobile network witnesses tremendous growths in the data traffic, the amount of user equipment, the diversity of applications, and the polymorphism of service scenarios. Cisco has predicted a seven-fold increase in the global mobile data traffic between 2016 and 2021, where the vast majority of traffic will be generated by portable devices [1]. Meanwhile, a report released by the Fifth Generation Infrastructure Public Private Partnership (5GPPP) addresses some expectations for the next generation of mobile networks, including a support to more than 10^4 devices per square kilometer, a high data rate up to 1 Gbps, and an ultra low transmission delay between 1 to 10 ms [2]. In order to meet end-user requirements beyond 2020 and optimize legacy networks upon future demands, mobile network operators (MNOs) have to find efficient solutions to enhance the Quality of Service (QoS), increase the spectrum efficiency, and maintain a healthy revenue while reducing the Capital Expenditure (CAPEX) and Operational Expenditure (OPEX).

To counter the traffic growth, build cost efficient networks and provide enhanced QoS to massive end-users, numerous architectures and technologies have been proposed for the Fifth Generation (5G) mobile networks, especially in the domain of Radio Access Network (RAN). We can generally categorize them into the four categories as follows:

1) Improving the spectrum efficiency for higher data capacity by deploying advanced transmission techniques such as massive Multi-Input Multi-Output (massive MIMO), Millimeter Wave (mmWave)

The associate editor coordinating the review of this manuscript and approving it for publication was Jafar A. Alzubi.

transmission, and beamforming [3]. There have been significant progresses in the development and deployment of these technologies. However, they are still constrained by major technical challenges, including implementation complexity, Radio Frequency (RF) interference, environmental obstacles, and antenna correlation [4].

2) Deploying small cells (pico cells, femto cells, and micro cells) combined with macro cells upon the existing network infrastructure. This approach has been used by Long Term Evolution - Advanced (LTE-A) networks as well, but the heterogeneity level in 5G RAN is significantly higher in comparison to that in legacy RAN architectures. Moreover, the deployment of small cells leads to increases in the power consumption, the CAPEX/OPEX, a variety of interference, and the handover frequency [5], [6].
3) Applying emerging technologies of Software Defined Network (SDN) and Network Function Virtualization (NFV) to softwarize the networks [7]. However, limitations to the deployment of both NFV and SDN widely exist in security, management, orchestration, isolation, resources allocation, dynamicity, flexibility, and scalability [8], [9].
4) Optimizing and reconstructing the network architecture – specifically the RAN architecture – by bringing computation, communication, processing and storage devices close to the edge of networks, so that end-users can access the data and services with low latency and high throughput [10]. However, this approach leads to increased CAPEX/OPEX as well as demands for new protocols and interfaces [11].

Out of the four aforementioned categories, in this article we survey and compare a series of state-of-the-art literature on various RAN architectures and key enabling RATs towards 5G mobile communication system.

A. REVIEW OF RELATED WORKS

Several survey papers on 5G communication system have been published so far. In [7], the authors address ten key enabling technologies towards 5G communication, including wireless SDN, NFV, mmWave, massive MIMO, network ultra-densification, big data & mobile cloud computing, scalable Internet of Things (IoT), device-to-device connectivity with high mobility, green communications, and new radio access techniques. They have also provided the challenges and limits for every presented technology. The authors of [11] have provided an overview of 5G network architectures. Along with this, major enabling technologies for 5G such as ultra-dense networks, multi-radio access technology association, interference management, spectrum sharing with cognitive radio, full duplex radios, etc. are surveyed. In [12], an overview is provided on 5G research, standardization trials, and deployment challenges. In addition, the authors identify key enabling 5G technologies, evaluate their strengths and weaknesses, and outline their research challenges. The authors of [13] investigate the flexibility and adaptability requirements that 5G shall achieve with radio access technologies and emerging system-level techniques. The work [14] overviews network architectures and promising techniques towards 5G, highlights the state-of-the-art, and further studies the implementation issues related to addressed techniques. In [15], the authors provide an exhaustive review of architectural changes associated with the RAN design such as air interfaces, smart antennas, cloud and heterogeneous RAN. The paper also studies key enabling physical layer technologies and applications towards 5G. The authors of [16] investigate advantages, applications, proposed architectures, implementation issues, real demonstrations, testbeds, emerging technologies, and research challenges towards 5G networks. The paper also presents a comparative study of proposed 5G architectures in the perspectives of energy efficiency and network hierarchy. In [17], the ongoing researching works on key enabling technologies for 5G are comprehensively reviewed. The authors have further addressed the research progresses of the key technologies and service models for 5G mobile systems and networks.

B. CONTRIBUTIONS

In contrast to the existing surveys introduced in subsection I-A, this article presents a more comprehensive literature review and comparative analysis of various RAN architectures towards 5G mobile systems. We also provide an extensive discussion on the key enabling technologies of 5G RAN architectures. Based on this, we name the contributions of this article as follows:

- We provide a detailed discussion on 5G communication system, service categories, proposed system architecture, advantages, applications, and implementation issues. Along with this, we also address the performance requirements, deployment of 5G in certain countries, and enabling of various vertical industries through network slicing.
- We present a comprehensive review of the evolution of historical and current RAN architectures. Based on this, various RAN implementations reported in the literature are comprehensively discussed. We also present the state-of-the-art, the ongoing standardization activities, and the motivation of reconstructing & redesigning legacy RAN architectures with respect to the requirements of next generation mobile networks.
- We address an in-depth study and a literature review of Cloud-RAN (C-RAN), Heterogeneous Cloud RAN (H-CRAN), Virtualzied Cloud RAN (V-CRAN), and Fog RAN (F-RAN). Moreover, we attempt to explore the aforementioned RAN architectures from various perspectives such as energy consumption, security, CAPEX/OPEX, performance, spectrum, mobility, resource allocation, and system architecture.
- As one of the key contributions of this paper, we provide an extensive comparative analysis of

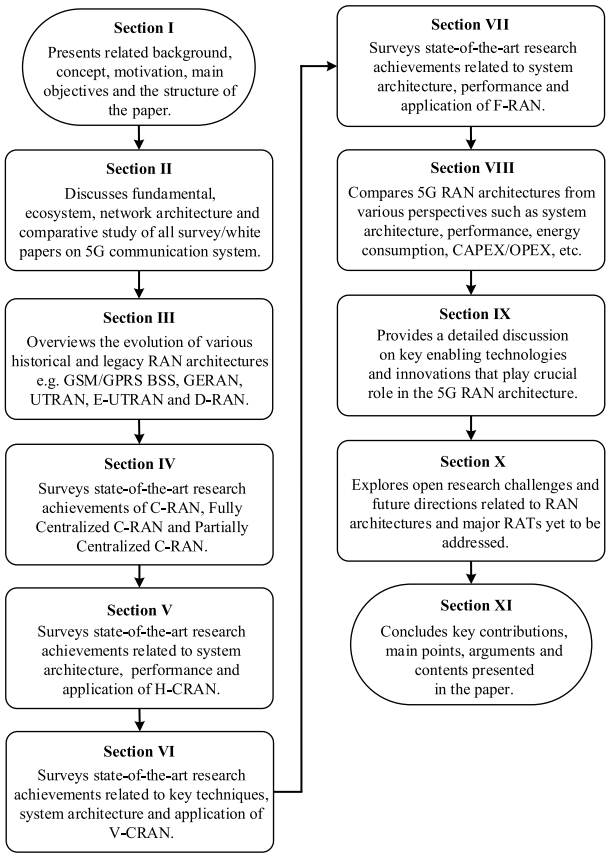

FIGURE 1. The condensed structure of this article.

C-/H-C/V-C/F-RANs, aiming to study these RAN architectures from various perspectives such as energy consumption, system architecture, CAPEX/OPEX, system level performance, etc.

- Moreover, we present an exclusive review and a comprehensive survey of the key enabling technologies for 5G RAN. Along with this, we explore the innovations in 5G RAN, which are expected to enhance spectrum efficiency, decrease CAPEX/OPEX, and fulfill the end-user expectations and of vertical industrial requirements.
- Last but not least, we address existing challenges and future research directions related to 5G RANs and RATs, expecting to motivate the community towards realistic solutions with new technical advances.

C. STRUCTURE OF THE PAPER

This paper is structured as illustrated in Fig. 1. In Section II, we generally introduce the 5G mobile communication system. Section III discusses the evolution of various traditional and legacy RAN architectures. The C-RAN, H-CRAN, V-CRAN and F-RAN are surveyed in Sections IV, V, VI and VII, respectively. We further compare all the named RAN architectures from various perspectives in Section VIII. Section IX gives a detailed description of major RATs and innovations, which play a key role in terms of enhancing performance, decreasing CAPEX/OPEX and simplifying operations in the 5G RAN architecture. We then give some outlooks to the future research directions in Section X, before closing this article with our conclusions in Section XI. Note that some acronyms frequently used in this article are listed with their definitions in Table 1 for the ease of reference.

II. THE 5G MOBILE COMMUNICATION

The Second Generation (2G) of mobile communication systems focused on voice service, while the focus of Third Generation (3G) and Fourth Generation (4G) systems shifted to data and mobile broadband services. Moreover, the mobile broadband service will be further pushed forward by the deployment of 5G systems to support an assortment of emerging uses cases, which may involve with ultra-high volume of data traffic, massive number of connections and high user mobility.

The 5G systems are expected to cope with future demands for service and business beyond the horizon of 2020, play a vital role in enabling of new innovations, and prove a significant boon to the economic output. Many countries and regions, such as US, EU, China, Japan, UK, South Korea, etc., are actively participating the race towards 5G, attempting to establish their technical and economic leaderships in the next decade. For example, Korea Telecom in South Korea launched a mmWave-based 5G communication system at the 2018 winter Olympics, while Japanese operators have planned to demonstrate their 5G system in 2020 during the summer Olympics in Tokyo [18].

The European Commission (EC) has also launched several 5G research and development projects in two innovation and research funding programs, namely the Seventh Framework Program for Research and Technological Development (FP7) and the Horizon 2020 (H2020). In FP7, the projects aim to explore the requirements, functionalities and architectures for beyond 4G and 5G communication networks. Whereas, in H2020, the projects are focusing more deep into the network softwarization and/or virtualization, network slicing, common test-beds, applications, spectrum, access network, security, etc. The EC has placed the digitization of various vertical industries such as automotive, healthcare, agriculture, etc. on the core of its 5G action plan published in 2016 [19]. Moreover, the commission is closely collaborating with the EU member states in order to ensure that its 5G action plan is efficiently and coordinately implemented step-by-step – aiming to make the dream of 5G into a reality for all citizens and verticals by the end of this decade [19].

On the other hand, various telecommunication vendors, operators and research institutions have also been contributing to the research and development of various aspects of 5G systems such as key technologies and enablers, spectrum, requirements, network architecture, use-cases, applications, etc. They have reported the progress of their contributions in a number of white-papers including [20]–[38]. These white-papers are listed in an ascending order along with a summary of their key contributions in Table 2.

Moreover, standardization organizations around the world have already initiated the standardization process for a global

TABLE 1. List of important acronyms.

Acronym	Definition	Acronym	Definition
5G	Fifth Generation	NF	Network Function
5GPPP	5G Infrastructure Public Private Partnership	NFV	Network Function Virtualization
BBU	Baseband Unit	NGMN	Next Generation Mobile Networks
BSS	Base Station Subsystem	NIC	Network Interface Controller
BTS	Base Transceiver Station	NLOS	Non-line of Sight
CAPEX	Capital Expenditure	NOMA	Non-Orthogonal Multiple Access
CDMA	Code Division Multiple Access	OFDM	Orthogonal Frequency Division Multiplexing
C-RAN	Cloud Radio Access Network	OLT	Optical Line Terminal
C-MTC	Critical-Machine Type Communication	ONU	Optical Network Unit
CN	Core Network	OPEX	Operational Expenditure
D2D	Device to Device	PCU	Packet Control Unit
D-RAN	Distributed Radio Access Network	PON	Passive Optical Network
DU	Digital Unit/Data Unit	QoS	Quality of Service
eMBB	enhanced Mobile Broadband	RA	Random Access
EPC	Evolved Packet Core	RAN	Radio Access Network
F-AP	Fog Access Point	RAT	Radio Access Technology
FDMA	Frequency Division Multiple Access	REC	Radio Equipment Controller
F-RAN	Fog Radio Access Network	RNC	Radio Network Controller
F-UE	Fog User Equipment	RRH	Remote Radio Head
GPRS	General Packet Radio Service	RRM	Radio Resource Management
GSM	Global System for Mobile Communication	RRU	Remote Radio Unit
H-CRAN	Heterogeneous Cloud RAN	SDN	Software Defined Networking
HPN	High Power Node	SDR	Software Defined Radio
HSPDA	High Speed Packet Downlink Access	SEG	Security Expert Group
HTC	Human-Type Communication	S-GW	Serving Gateway
ICN	Information Centric Networking	SS7	System No. 7
IoE	Internet of Everything	TCO	Total Cost of Ownership
IoT	Internet of Things	TCP	Transmission Control Protocol
ISG	Industry Specification Group	TDD	Time-Division Duplex
LC	Line Card	TDM	Time Division Multiplexing
LoS	Line of Sight	TDMA	Time Division Multiple Access
LTE	Long Term Evolution	UDN	Ultra-Dense Networking
LTE-A	Long Term Evolution-Advanced	UMTS	Universal Mobile Telecommunication System
MAC	Media Access Control	URLLC	Ultra Reliable Low-Latency Communications
MEC	Multi-access Edge Computing	UTRAN	UMTS Terrestrial Radio Access Network
MIMO	Multi-Input Multi-Output	VBS	Virtualized Base Station
MME	Mobility Management Entity	V-CRAN	Virtualized Cloud RAN
mMTC	massive Machine-Type Communications	VNF	Virtual Network Function
mmWave	Millimeter Wave	WCDMA	Wideband Code Division Multiple Access
MSC	Mobile Switching Center	WNC	Wireless Network Cloud

5G radio interface. The 3rd Generation Partnership Project (3GPP), as one of the well-know organizations, has been actively involved in the standardization process of communication systems since the end of last century. Their standards of 5G have been proposed for the first time in their Release (Rel.) 15 , with further progresses in the undergoing Rel. 16. The 3GPP 5G standards are expected to become available for commercial deployment and service delivery between 2020 and 2030.

A. THE 5G AMBITIONS

5G is supposed to provide significant advances in all aspects of performance, including a 1000-fold growth in system capacity, an enhanced connectivity to at least 100 billion devices, 10 Gbps maximum and 100 Mbps average individual user experience, prolonged battery life with 1000-fold lower energy consumption per bit, a 90% reduction in network energy usage, support to 500 km/h mobility for high speed users (e.g. high speed trains), a 3-fold increase in spectrum

TABLE 2. A summary of individual white papers focusing on 5G communication system.

Reference	Key Contributions
[20]	Overviews key enabling technologies, requirements, challenges, time-line, architecture and characteristics.
[21]	Addresses concept, key technologies, requirements, architecture, characteristics, time-line and end-to-end network slicing.
[22]	Discusses motivation, concept, key enablers, architecture, time-line, requirements and technical scenarios.
[23]	Provides a detailed discussion on spectrum for 5G and more specifically on harmonization of spectrum and spectrum ranges.
[24]	Investigates architecture, concept, key technologies, technical scenarios, time-line and requirements.
[25]	Covers key enabling technologies, requirements, key performance indicators and architecture.
[26]	Presents concept, use-cases, services, requirements, network slicing, architecture and applications in vertical industries.
[27]	Sheds light upon requirements, vision, characteristics, architecture, technology, spectrum and intellectual property rights.
[28]	Addresses requirements, vision, architecture, applications, key technologies, services and spectrum.
[29]	Investigates vision, requirements, architecture, use-cases and radio access network architecture.
[30]	Provides a detailed discussion on concept, key technologies, use-cases, cost implications, spectrum and radio access technologies.
[31]	Discusses specifications, requirements, characteristics, use-cases, global 5G development and research challenges.
[32]	Addresses requirements, future global mobile data forecast, characteristics, applications and distributed antenna system.
[33]	Overviews concept, requirements, characteristics, key technologies, activities and technical components.
[34]	Presents concept, requirements, implications, characteristics, use-cases and enabling applications of 5G in vertical industries.
[35]	Sheds light upon concept, requirements, key technologies, air-interface, capabilities and applications.
[36]	Provides a detailed discussion on concept, characteristics, requirements, applications and key technologies.
[37]	Discusses concept, requirements, characteristics, spectrum, radio interface, next generation of radios and applications.
[38]	Investigates requirements, characteristics, applications, key enabling technologies and deployment.

efficiency, perception of 99.99% availability, 100% coverage, and latency from 1 to 10 milliseconds [20], [39]. In order to achieve these goals and fulfill the requirements of end-users and industry beyond 2020, drastic improvements and disruptive innovations of emerging mobile network architecture design are needed, on both physical and upper layers. This improvement and reconstruction of network architecture do not only increase system level performance but also enhance energy efficiency and decrease CAPEX and OPEX.

Enhancing energy efficiency is a key pillar in the development, standardization and deployment of 5G mobile communication system. With full deployment of 5G mobile networks, there will be millions of Base Stations (BSs) and billions of connected devices around the globe that need for energy-efficient operations and systems. For the time being, the Information and Communication Technology (ICT) industry and systems are responsible of 5% of world's Carbon Dioxide (CO_2) emissions [40]. This level of emission is increasing globally along with the increases of connected devices, networks, and data/VoIP traffic. Moreover, it is predicted that 75% of ICT sector will be wireless by 2020, therefore, energy consumption of the ICT industry is considered as one of the main global environmental concerns. In order to reduce energy consumption and to decrease the CO_2 emissions globally, new approaches to wireless communication networks are demanded [41].

Every single 5G technology has its own impact on both the CAPEX and the OPEX of a mobile operator. Some of the technologies decrease the costs such as virtualization of network functions, expectedly by about 30% [42], while

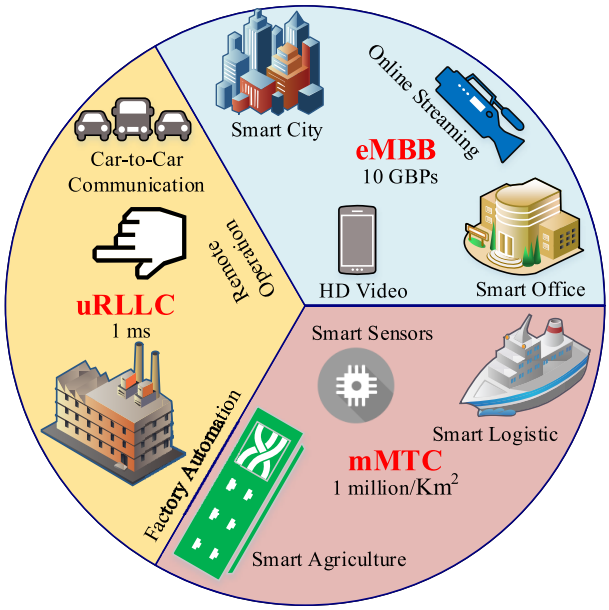

FIGURE 2. 5G mobile communication service categories.

some will increase the costs such as the limited propagation characteristics of high-frequency spectrum. The study [43] shows that mobile operators spend 60-80% of CAPEX on RAN technologies. On the other hand, China Mobile Research Institute (CMRI) claims that by adopting C-RAN, a 15%-reduction in CAPEX and a 50%-reduction in OPEX can be achieved [44]. Therefore, it is important to consider the appropriate deployment of each of the use-cases in order to reduce the Total Cost of Ownership (TCO) of an operator.

B. THE 5G SERVICE CATEGORIES

The development of 5G communication systems focuses on three fundamental issues, namely increased capacity, massive connectivity and diverse set of services [20]. As shown in Fig. 2, the International Telecommunication Union (ITU) has classified all 5G communication services into three categories: enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC) and Ultra-Reliable Low-Latency Communications (URLLC). The eMBB is expected to meet the ordinary end-user demands, i.e., higher bandwidth for Internet access suitable for web browsing, virtual reality, streaming of High Definition (HD) videos, etc. The mMTC aims at services for massive amount of connected devices and machines, e.g., smart city, smart agriculture, sensor networks, smart metering, etc. The URLLC tends to serve latency-critical applications such as autonomous driving, factory automation and remote surgery, which usually require sub-millisecond latency with packet loss rate below 10^{-5} [21].

The aforementioned service categories of 5G systems shall also provide support to a variety of vertical industries [45], including health-care, manufacturing, automotive, logistic, energy, environment, construction, etc., as shown in Fig. 3. These industries call for various use-cases with different QoS requirements, which eliminates the traditional one-size-fits-all network architectural approach from the utilization of 5G and beyond communication technologies. In order to efficiently accommodate vertical use-cases along with increased user demands over the same network infrastructure, 5G requires an architectural optimization and reconstruction with respect to the current deployment. The network will be logically sliced into different virtual networks, a.k.a. the network slices, in order to meet diversified service requirements and provide flexible support to various application scenarios [46]. In addition, the technology of network slicing also allows the operators to partition their networks in a structured, elastic, scalable and automated manner.

C. THE 5G KEY ENABLERS

The key enabling technologies of 5G communication system are from both wireless and networking areas. In the field of networking, MEC, SDN, NFV and network slicing are considered to be the main key enabling technologies of 5G. In the field of wireless, massive MIMO, Ultra-Dense Networking (UDN), novel multiple access, and all-spectrum access are of 5G's interest. In addition, technologies such as Device to Device (D2D) communication, flexible duplex, full duplex, mmWave, device-centric architectures, etc., are also considered as 5G key enabling technologies [7], [11], [47]. We provide a detailed discussion on some of these key enabling technologies in Sec. IX.

D. THE 5G SPECTRUM

The 5G mobile communication system needs new spectrum (both licensed and unlicensed) in order to fulfill the requirements of mobile phone users and vertical industries. Worldwide, there are significant on-going efforts to identify suitable spectrum, which also includes bands that can be used in as many countries as possible to enable global roaming. Extremely high-frequency spectrum is considered as one of the most prominent candidates. According to [42], the frequency spectrum for 5G is broken into three sections: low-band, mid-band and high-band. The low-band spectrum is below 1 GHz, the mid-band spectrum spans from 1 GHz to 6 GHz and the high-band spectrum is above 24 GHz, which is often called millimeter waves band [42]. Each band has its unique characteristics that suit for certain deployment scenarios. More specifically, the low-band has good propagation characteristics, which is beneficial for covering a large area. However, it has limited capacity due to the lack of bandwidth. The mid-band provides a coverage that is feasible for urban deployment, with an increased capacity. The high-band has the most limited coverage, while providing a high capacity with its richness of available spectrum [23].

E. THE 5G MOBILE NETWORK ARCHITECTURE

In order to meet 5G requirements, a new architecture is required to accomplish a total network revolution. The 5G network architecture is consisted of a simplified but efficient core network with control and forwarding

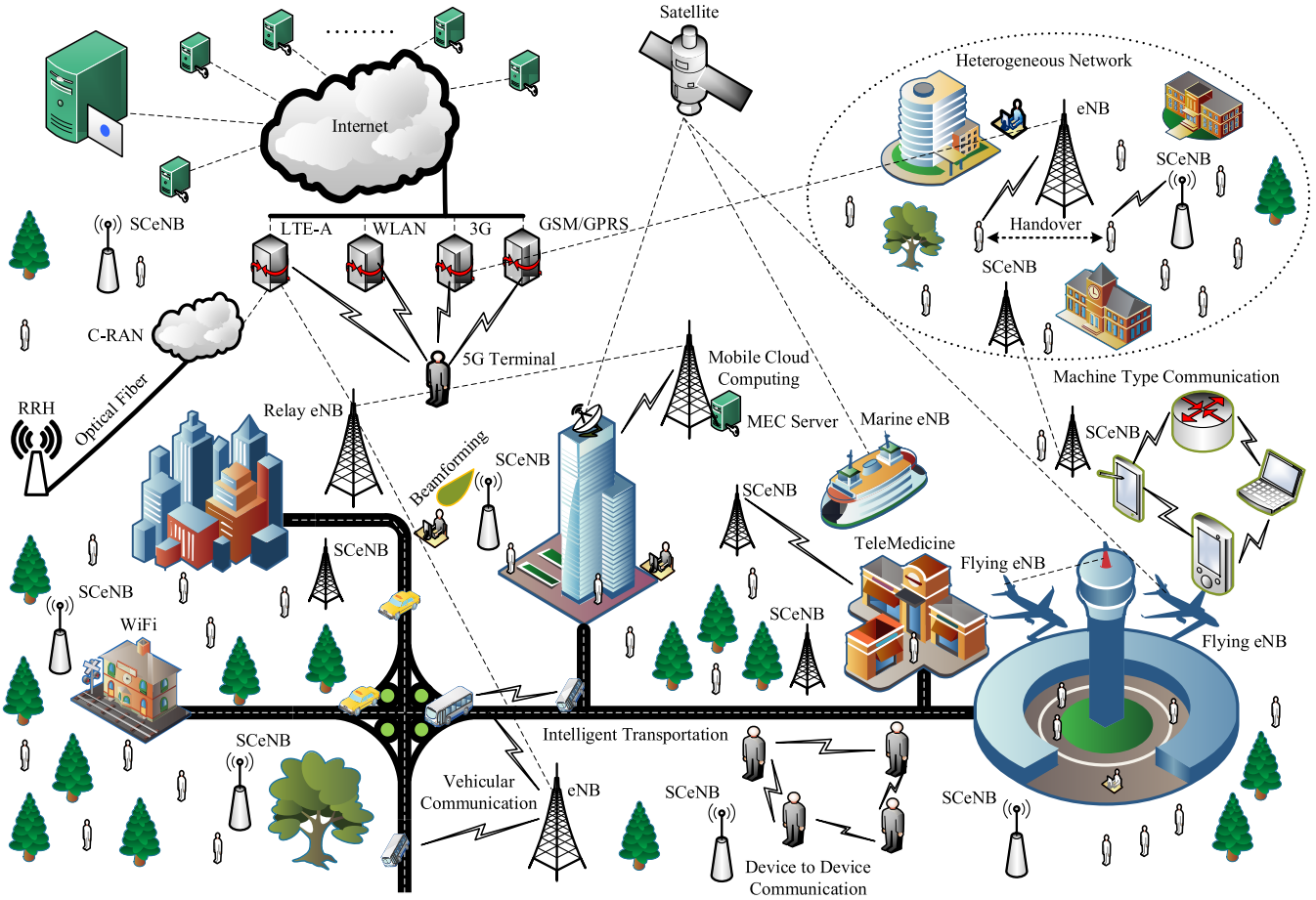

FIGURE 3. 5G mobile communication ecosystem.

functions, and a high-performance access network. As illustrated in Fig. 4, the 5G logical network architecture is composed of Access Plane, Control Plane and Forwarding Plane [22]. The access plane is consisted of various types of BSs and access devices. There is an enhanced interaction and rich networking topology among BSs and wireless devices, which leads to flexible access cooperative control and higher resource utilization. The control plane is in charge of generating of global control strategy for the entire network. The forwarding plane is responsible for forwarding of traffic from all network devices and resources. The efficiency and flexibility of data forwarding can be achieved through schedule policies generated by the unified control plane. From the infrastructural point of view, as Fig. 4 further shows, a 5G network is composed of Access Network, Metropolitan Area Network (a.k.a. Aggregation Network) and Backbone Network. The control functions can be categorized into core network control functions and access network control functions. The core network control functions are deployed centrally within the aggregation and backbone networks, while the access network control functions are deployed at the edge of the mobile network or integrated into the BS in order to provide supports to low latency and high reliability services.

So far, we have thoroughly discussed the characteristics, requirements, applications, network architecture and other crucial aspects of 5G mobile communication. In coming sections, we are going to take a detailed look to various RAN architectures proposed for 5G mobile network. The interested readers will learn that 5G access network is a multi-layer heterogeneous network in order to satisfy various use-case scenarios and applications. It is a comprehensive domain of 5G network consisting of various types of RAN architectures and a combination of macrocell, microcell, femtocell, picocell, and unified multi-access technologies, which together improve cooperative processing efficiency of cell edge and utilize backhaul/fronthaul resources. Moreover, we will also take a look to new RATs for 5G communication system. The 5G RATs not only enhance performance in the RAN, however, they also introduce more choices so that operator can utilize them in various scenarios. But, first we are going to thoroughly discuss the historical and legacy RAN architectures of mobile communication networks in the next section.

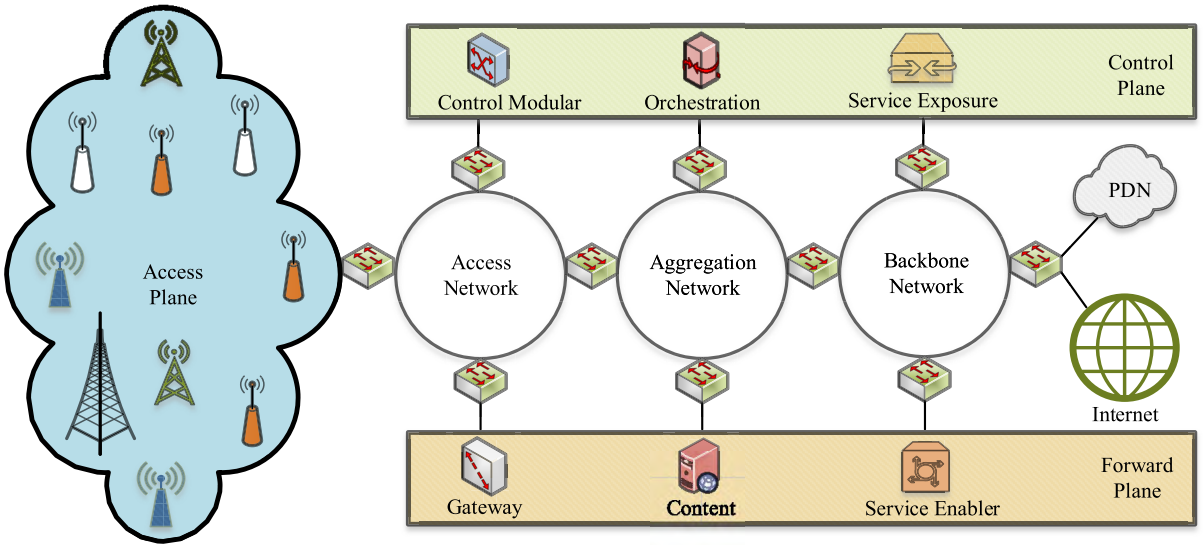

FIGURE 4. 5G mobile network architecture.

III. THE EVOLUTION OF RAN ARCHITECTURES

Along the history of mobile network evolution, every architectural innovation is driven by new requirements of mobile communication services that cannot be satisfied by legacy systems, which have inherited shortages in their outdated architecture. In order to address the RAN architecture for 5G mobile systems, it is essential to track back the development of previous and current RANs. In this section, we are going to discuss the concept and architectural evolution of RAN in various historical and legacy systems.

A. THE BASE STATION SUBSYSTEM (BSS)

The BSS is the core of 2G RAN architecture [48], which is standardized as part of the Global System for Mobile Communication (GSM). The main goals of the BSS are to provide network coverage for a desired area and to perform radio and mobility functions. The coverage area of every BSS spans over multiple small areas, which are called cells [49]. Each cell is served by at least one fixed-location transceiver or the Base Transceiver Station (BTS). In regions such like Europe, the term cell is also known as sector. In contrast, in the USA a cell refers to the coverage area of a single BTS that consists of a group of sectors, and is typically divided into three sectors [50]. A set of cells served by the same Base Station Controller (BSC) is called a cluster. The size, shape, capacity and network coverage of a cell depend on both the user density and the topography. A cellular network enables a large number of Mobile Stations (MSs) in its coverage area to communicate with each other, with MSs of other mobile operators and with fixed landline telephones.

In order to support a large number of MSs within a limited spectrum, the concept of frequency reuse was developed. In this concept, the same frequency can be reused by multiple BSs that are sufficiently spaced apart (geographically/physically) [49]. Radio channels are distributed over a cell area in such a way that co-channel interference is insignificant. Assume a cellular system with S available duplex channels and let N be the number of cells in a cluster, where each cell uniformly allocates K duplex channels. Then, the disjoint channel groups would be $S = KN$. If the cluster pattern is repeated M times within a region, the total number of duplex channels will be $T = MS = KMN$. Hence, we can conclude that by deploying of frequency reuse, it achieves a capacity gain proportional to the number of times a cluster pattern is repeated. As shown in Fig. 5 (RAN 1), the available spectrum is divided into seven frequency bands (F1-F7) and each band is allocated to a cell. The allocated spectrum can be reused elsewhere considering a sufficient inter-cell distance, e.g. RAN 2.

As shown in Fig. 5 (RAN 1), the BSS consists of the BTS, the BSC, the Air-interface, the Abis-interface and the A-interface. In GSM networks, BTS is the first element with direct wireless connection to MSs. It consists of antennas and radio front-end hardware in order to communicate with MSs over radio link. The BSC is responsible for the mobility management and radio resource management of all BTSs their connected MSs. A typical BSS includes tens of BSCs and hundreds of BTSs. These nodes and the entire BSS architecture bridge the gap between millions of MSs and the GSM Core Network (CN). The air-interface connects the MSs and the BTS, and allows MSs to communicate with each other. The abis-interface (typically an E1 link) is used to connect BTSs to the BSC. It is a channelized Time Division Multiplexing (TDM) link, where each user connection requires 16 Kbps or 8 Kbps depending on the modulation scheme [48]. The A-interface (combination of multiple E1s) is used to connect BSCs to the CN.

B. THE GENERAL PACKET RADIO SERVICES BSS (GPRS BSS)

The importance of Internet for portable devices increased in mid-1990s. In 1997, the European Telecommunications

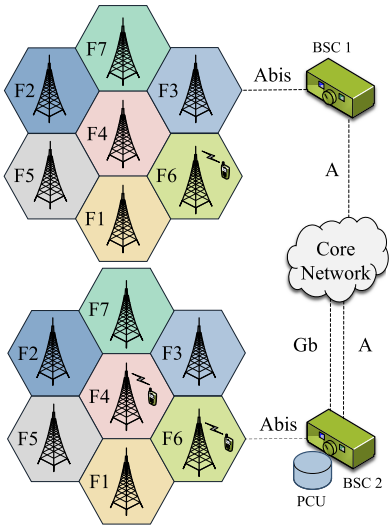

FIGURE 5. GSM/GPRS RAN architecture.

Standards Institute (ETSI) certified the specifications of GPRS, which was fully accomplished in 1999 [51]. The GPRS technology started its integration into the cellular networks in 1999, but became commercially available in 2001. Adding GPRS functionalities to the GSM network has actually brought necessary changes in the network architecture, particularly in the BSS.

The GPRS uses packet-switching, where the available capacity is shared among many users, so that the waste of bandwidth is decreased to a low level. In comparison to circuit-switching, packet-switching is more efficient with bandwidth utilization [51]. Packet-switching is considered as a turning point in the history of cellular communication, which opened the door for research and development of 3G and beyond technologies. A comparison between packet switching and circuit switching techniques along with their pros and cons is thoroughly described in [51], [52].

The GPRS and GSM operate alongside one another using the same BSS architecture. As GPRS provides a move from circuit to packet switching, it calls for an upgrade of the BSS architecture. As shown in Fig. 5 (RAN 2), a new element, the Packet Control Unit (PCU), was added to the BSS. Meanwhile, existing elements and interfaces were also modified. The GPRS/GSM standards allow the PCU to be positioned at various locations within the network, i.e., close to BTS, BSC or CN. The most common deployment position of the PCU is close to the BSC as shown in Fig. 5 (RAN 2) [51]. In addition to the PCU, at least one separate E1 link from BSC to BTS is needed to support GPRS services, e.g. web browsing and file downloading. This link does not change the basic architecture of BTS or the protocol requirements of an E1 [51].

C. THE GSM/ENHANCED DATA RATE FOR GSM EVOLUTION (EDGE) RAN (GERAN)

The GERAN is the access network for EDGE, which is specified in GSM Phase 2+ Rel. 98 [53], and further improved

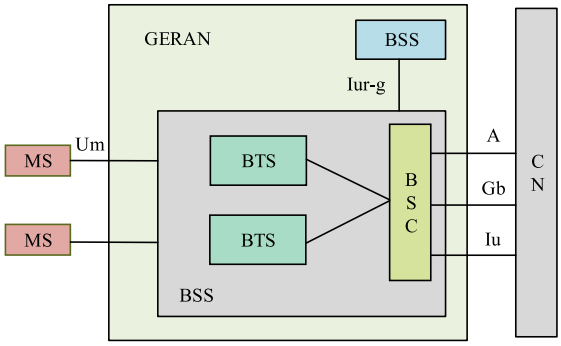

FIGURE 6. The GERAN architecture.

in 3GPP Rel. 5 and Rel. 6 later on [53]. The Rel. 5 specifies an interface called Iu, which connects GERAN to 3G CN. This leads to a new architecture for GERAN and significant modifications to its radio protocols [53], as illustrated in Fig. 6. The Rel. 6 specifies major enhancements on the physical layer for GERAN.

The main motivation of the GERAN implementation is to improve the data rate of GSM/EDGE network and to increase end-users' satisfaction. More specifically, the EDGE aims to increase the data transmission rate per radio time-slot by enhancing the Gaussian Minimum-Shift Keying (GMSK) modulation used in GSM/GPRS networks. The radio interface in the GERAN uses GMSK with 8 Phase Shift Keying (8-PSK) [54], which transmits three bits per symbol (instead of one bit per symbol as in GSM/GPRS networks). This evolution in modulation scheme increases the peak bit rate from approximately 20 Kbps to around 60 Kbps per timeslot.

The general architecture of GERAN is presented in Fig. 6. The Um interface is used to connect the MS with the BTS of the GERAN, the Gb interface is used in GSM/GPRS to connect Serving GPRS Support Node (SGSN) and BSS, while the A interface is used to connect BSS and 2G Mobile Switching Center (MSC) which are supported by GERAN. There are two new interfaces in GERAN, the Iu and the Iur-g. The former, as aforementioned, connects GERAN with the CN. The latter connects GERAN with RANs of other architectures, such as GSM/GPRS or Universal Mobile Telecommunications System (UMTS) RANs.

D. THE UMTS TERRESTRIAL RADIO ACCESS NETWORK (UTRAN)

The UTRAN for UMTS was standardized for the first time in 3GPP UMTS Rel. 99 at the end of 1999 [55]. The UTRAN is built up on existing standards, therefore is strongly influenced by previous RAN architectures. The following release of UMTS specifications was originally known as Rel. 2000, and then decomposed into two 3GPP specification releases, namely Rel. 4 and Rel. 5. The former specifies all optional changes in the circuit-switched side of UMTS CN. Whereas the latter defines the IP multimedia subsystem, High Speed Packet Downlink Access (HSDPA) and other multimedia mechanisms [56], [57].

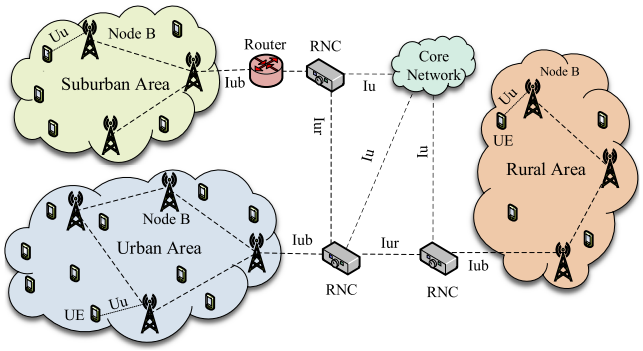

FIGURE 7. The UTRAN architecture.

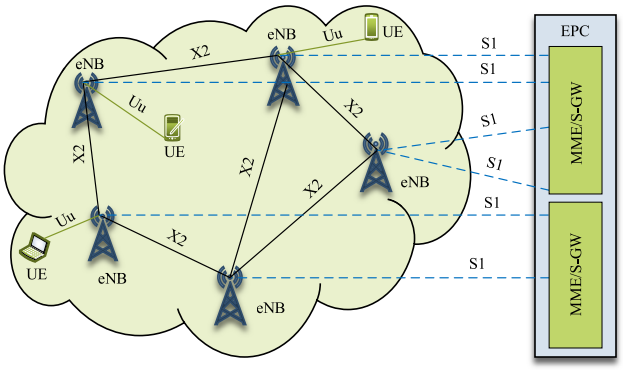

FIGURE 8. The E-UTRAN architecture.

We provide an example of the UTRAN in Fig. 7. As illustrated, the UTRAN consists of one or more Radio Network Subsystems (RNSs), each comprises of at least one Radio Network Controller (RNC) and a number of BSs [58], [59]. The BS and air interface in the UTRAN are called Node B and Uu, respectively. The Uu interface uses Wideband Code Division Multiple Access (WCDMA), which is based on Direct Sequence Spread Spectrum (DSSS) and Code Division Multiple Access (CDMA), in order to achieve higher speed and support more simultaneous user links in comparison with Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA). The RNC communicates with Node B and the CN over two transmission links, known as the Iub interface and the Iu interface, respectively. There are two types of Iu interfaces: one for circuit-switching CNs and the other for packet-switching CNs. The RNC is the central element of the UTRAN, which is responsible for the mobility management of UE and Radio Resource Management (RRM) of all attached cells. Moreover, the RNC is also responsible for the setup, release and maintenance of Radio Barriers (RBs) [58].

The density of end-users and BSs vary from one location to another (see Fig. 7), therefore, the design requirements of UTRAN also strongly dependents on the terrain. The main differences between various terrains such like urban, suburban, and rural areas are the population density and the required capacity. In rural areas, due to the low population density, a lower number of Node B stations is required in order to provide network coverage. In suburban areas, the density of the UEs is higher than that in rural areas. Therefore, it requires more Node B stations to achieve the same coverage. In urban areas, the number of UEs is undoubtedly higher and requires a dense deployment of Node B stations. It is a common strategy to densify BSs in the RAN so as to improve the data rate in user-dense areas. However, this leads to further challenges in networks, such as stronger interference and frequent handovers. In order to mitigate interference between adjacent Node B stations, BSs shall be installed sufficiently distant apart from each other. A typical density of BSs in UTRAN ranges between 4 to 5 BSs per km^2 [60].

E. THE EVOLVED UTRAN (E-UTRAN)

The E-UTRAN was standardized for the first time along with LTE in Rel. 8 and Rel. 9 [61], [62]. Unlike the UTRAN and BSS, the E-UTRAN consists of no centralized controller but only base stations (eNode B), as illustrated in Fig. 8. Therefore, E-UTRAN is considered as a flat RAN in comparison to previous RANs [63]. The eNodeBs are interconnected with each other through X2 interface, and to the Evolved Packet Core (EPC) through S1 interface. Additionally, every eNodeB is connected to the Mobility Management Entity (MME) and the Serving Gateway (S-GW) through S1-MME and S1-U interfaces, respectively. The interface that connects eNobeB with the UEs is known as LTE-Uu [64].

Unlike previous RANs, the E-UTRAN integrates all functions including RRM, header comparison, security, etc., into the eNodeBs, which leads to a reduced latency and an improved efficiency [63]. In LTE, multiple EPC nodes, i.e. MME/S-GW, serve a single eNodeB through the S1 interface. This scheme provides a possibility for load sharing and eliminates the risk of single-point failure for the EPC nodes.

The Uu interface uses two different techniques to improve user experience for broadband data communications. The downlink is based on Orthogonal Frequency Division Multiplexing (OFDM) waveform and the uplink is based on Single-Carrier Frequency Division Multiplexing (SC-FDM) waveform. The S1 interface is splitted into a control plane and a user plane. The protocol structure over S1 is completely IP-based, and has neither dependency nor legacy with Signaling System No. 7 (SS7) network as was the case with UMTS and GSM networks. The control plane is based on Stream Control Transmission Protocol/IP (SCTP/IP) stack, while the data plane is based on GPRS Tunneling Protocol/User Datagram Protocol 5/IO (GTP/UDP5/IP) stack [65]. The X2 interface is used to exchange two kinds of information, the mobility and load/interference. The intra-LTE mobility has the highest priority in E-UTRAN, and therefore triggers the X2 interface by default, unless there is no X2 interface between source and destination eNodeBs. The exchange of load information over the X2 interface can be further divided into two types: the load balancing process, in which the information is used to balance the load; and the interference

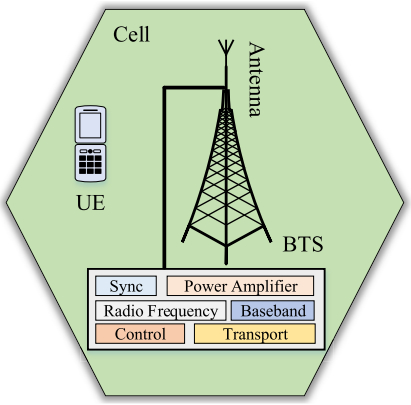

FIGURE 9. A traditional macro base station.

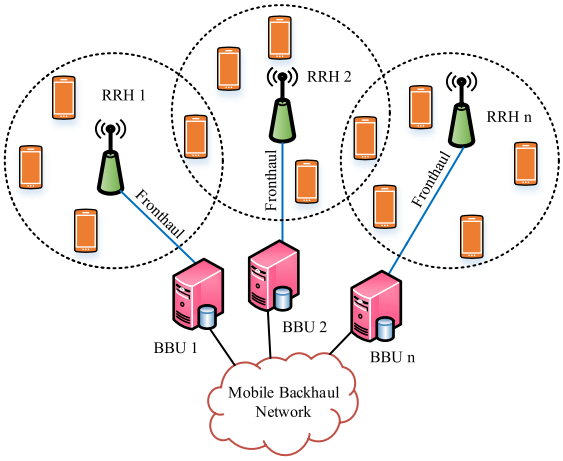

FIGURE 10. Traditional D-RAN architecture, where RRHs and BBUs are separated, however, every RRH is connected to its own dedicated BBU through fronthaul.

coordination process, in which the information is used to optimize some RRM processes. Protocol stacks of both user and data planes over the X2 interface are the same as those of S1, except for that X2-AP is substituted with S1-AP. This leads to simplify the data forwarding procedure [63].

LTE-A is standardized in Rel. 10 [66]. The main features of Rel. 10 are increased data rates (3Gbps downlink and 1.5Gbps uplink), allowing the combination of up to five separate carriers to enable bandwidths up to 100MHz, higher order MIMO antenna configurations of up to 8*8 in/for downlink and in/for 4*4 uplink, relay nodes to support heterogeneous networks containing a wide variety of cell sizes, and enhanced Inter-Cell Interference Coordination (eICIC) to improve performance towards the cell edge [50]. Further improvements of the E-UTRAN were brought later in releases (Rel. 11 - Rel. 14), by supporting services such as Narrow Band Internet of Things (NB-IoT), mMTC and D2D communications.

F. DISTRIBUTED-RADIO ACCESS NETWORK (D-RAN)

In the RAN architecture of 2G mobile networks (BSS), all radio and baseband processing functions are integrated into the BSs. A traditional BS consists of two functional devices, the Digital Unit (DU) and the Radio Equipment Controller (REC). The DU is responsible for functions such as amplification, modulation and demodulation, frequency conversion, radio frequency filtering, digital-to-analog and analog-to-digital conversion, whereas the REC is responsible for baseband signal processing, controlling and managing of BSs, and interfacing with the RNC [67]. We provide an example of a traditional macro BS in Fig. 9, where the DU and REC functional electronics are placed at the base of the tower and connected using coaxial cable with the antenna located at the top of the tower.

As shown in Fig. 10, the radio and signal processing units of a traditional macro BS are separated from each other in UTRAN and E-UTRAN. The radio unit, which is set close to the 3G/4G macro BS is called Remote Radio Head (RRH) or Remote Radio Unit (RRU). The baseband signal processing unit, which is located in a convenient and easily accessible location, is called Baseband Unit (BBU) or Data Unit (DU). The BBU dynamically allocates network resources to its corresponding RRHs with respect to the network requirements [68]. The RRH directly communicates with the end-user and is confined only to the RF functions. This RAN architecture is called the D-RAN (see Fig. 10). Each RRH is interconnected to its corresponding BBU through a transport network using Common Protocol Radio Interface (CPRI) in order to transmit In-phase and Quadrature (IQ) signal. Both optical fiber and microwave can be deployed for the link between RRH and BBU, which is called the fronthaul. According to [69], the distance between BBU and RRH in a same network can reach up to 40 km, which leads to processing and propagation delays. Meanwhile, a white-paper published by EXPO suggests that the length of optical fiber connecting RRH and BBU should be limited to 15-25 km [70], which is therefore the recommended length limit of fronthaul links. The link between BBU and CN is called the backhaul. It carries user data, control and management data, and handover data that are exchanged between eNBs. For the backhaul link, optical fibers are commonly deployed in transport networks.

To the best of our knowledge [67], [69], [71], the D-RAN is an efficient RAN solution for 3G and 4G mobile networks. Nevertheless, it is neither scalable, nor efficient enough to deliver the high bandwidth, low latency and cost efficient services expected by 5G. Therefore, we do not consider D-RAN as a competitive RAN solution towards 5G in this study.

IV. CLOUD-RADIO ACCESS NETWORK

As we have discussed in subsection III-F, all radio and baseband processing functions in 2G RAN are integrated into the BS. In contrast, with the deployment of D-RAN in 3G and 4G mobile networks, radio and baseband processing functions are splitted into two separate nodes, namely RRH and BBU. As the amount of user-data has been increasing with regard to various QoS requirements, network operators

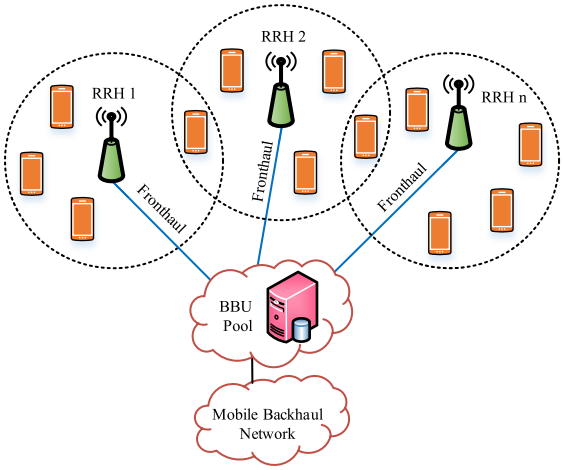

FIGURE 11. The C-RAN architecture, where RRHs and BBU are separated, however, all RRHs are connected to a shared and centralized baseband processing unit in a virtualized BBU pool through fronthaul.

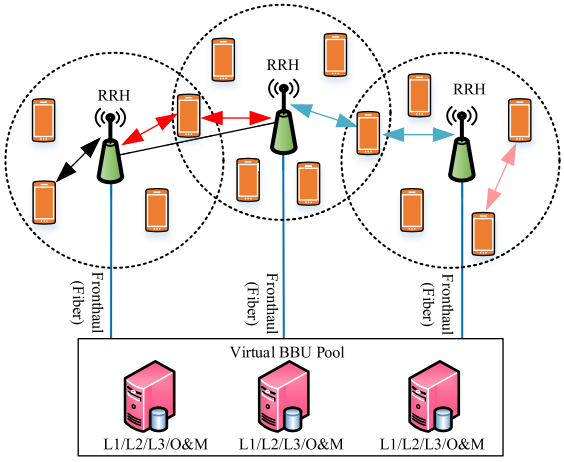

FIGURE 12. The fully centralized C-RAN solution, where all functions related to Layer 1, Layer 2, and Layer 3 are located in the BBU.

are forced to fulfill these requirements through centralization and cloudification of BBUs and their corresponding RRHs. This new centralized and cloudified RAN, where network resources are pooled in a centralized BBU pool, is known as C-RAN [72], [73].

C-RAN was initially proposed by IBM under the name Wireless Network Cloud (WNC) [74]. Later, it was described with further details in a CMRI white paper [73]. As depicted in Fig. 11, the main concept behind C-RAN is to separate all BBUs from their corresponding RRHs, and to pool them into a centralized, cloudified, shared, and virtualized BBU pool. Every RRH is connected through a fronthaul link to its corresponding BBU pool. Every BBU pool is able to support up to tens of RRHs, and connected through a backhaul link with the core network. The C-RAN architecture decreases the CAPEX and OPEX of MNO, reduces the energy consumption, increases the network scalability, simplifies network management and maintenance, improves the spectral efficiency and the network throughput, and facilitates load balancing.

The C-RAN embeds cloud computing into the RAN architecture of 5G [73]. It is initially based on two main tenets: the centralization and the virtualization of baseband processing. The centralization aims to optimize network performance, decrease energy consumption, and increase spectrum efficiency. Meanwhile, the virtualization of baseband processing aims to decrease the CAPEX and OPEX of 5G mobile networks.

A. TYPES OF C-RAN

With respect to the splitting of functions between RRH and BBU, C-RAN can be categorized into two types: Fully Centralized C-RAN and Partially Centralized C-RAN.

- In a fully centralized C-RAN, all functions related to Layer 1 (such as sampling, modulation, resource block mapping, antenna mapping, quantization, etc.), Layer 2 (such as transport-media access control), and Layer 3 (such as radio resource control) are located in the BBU (as shown in Fig. 12). Some key advances that fully centralized C-RAN brings to 5G mobile network are easy expansion of network coverage, easy upgrading of network capacity, support to multi-standard operation, enhancement of network resource sharing, and support to multi-cell collaborative signal processing. Despite all these advantages, fully centralized C-RAN faces two major challenges: the high bandwidth requirements, and the transmission of baseband I/Q signal between RRH and BBU [73]. Fully centralized C-RAN solution together with open platforms facilitate the development and the implementation of Software Defined Radio (SDR). This leads to an upgrade of the air interface through software, which furthermore helps upgrade RAN architecture more conveniently and provides support to multi-standard operation.
- In a partially centralized C-RAN, radio and baseband processing functions are integrated into the RRH, whereas all functions related to high layers are integrated into the BBU (as shown in Fig. 13). It specifically means that L1 related functions are located in RRH, whereas L2 and L3 related functions are located in BBU. Partially centralized C-RAN requires low transmission bandwidth between RRH and BBU, because the baseband processing is shifted from the BBU to the RRH. However, it also leads to some challenges, such as low flexibility in network upgrades and less convenience for multi-cell collaborative signal processing.

Both fully and partially centralized C-RANs have been studied and developed from various perspectives aiming to fulfill the requirements of 5G. The deployment of either of them depends on the network characteristics. In both cases, if the service provider is interested to expand the network coverage or split the cell in order to improve the capacity, it will be easy to deploy new RRHs and subsequently

TABLE 3. A summary of major works related to C-RAN in 5G mobile network.

Research Issues	Briefing	References
Energy	Decrease energy consumption of BSs, reduce transmit power, minimize downlink power, and increase energy efficiency through lowering number of air conditioning and site support equipment.	[73], [76]–[86]
Security	Potential attacks, security and privacy threats, security requirements, and proposed solutions are explored.	[87]–[90]
CAPEX/OPEX	CAPEX and OPEX are studied on cell site equipment, and transport network including both fronthaul and backhaul.	[69], [73], [91]–[94]
Performance	Studies have been carried out to explore mechanisms that increase overall performance including downlink/uplink data rates at cell edge.	[90], [95]–[98]
Spectrum	Spectrum efficiency, spectrum allocation, unlicensed spectrum, and sub-6-GHz-based licensed spectrum are studied.	[73], [75], [99]
Mobility	Mobile terminal handover, mobility management, mobility of high-speed end-users, and mobility performance are discussed.	[100]–[102]
Surveys	Cover architecture, security, NFV/SDN for C-RAN in 5G mobile networks.	[69], [75], [87], [103]

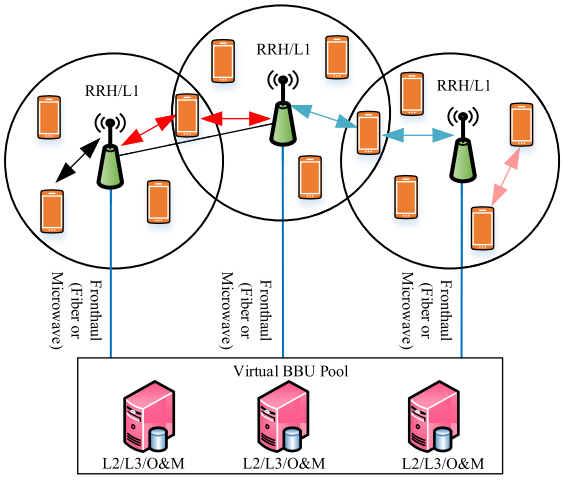

FIGURE 13. The partially centralized C-RAN solution, where Layer 1 functions are integrated in RRH, however, Layer 2 and Layer 3 related functions are located in BBU.

connects them with the BBU pool. On the contrary, if the service provider notices an increasing network load, it will only require to upgrade the hardware in the BBU pool so as to increase the processing capacity. It is worth noting that with the deployment of C-RAN, the static coupling between RRH and BBU is loosed. The RRH is no longer connected to a dedicated physical BBU. Instead, every RRH is connected to a virtual BS that is located in the BBU pool through a real-time virtualization technology.

B. STATE-OF-THE-ART

C-RAN has attracted a significant attention from research community. There have been numerous studies, which investigate C-RAN from various aspects. In the rest part of this subsection, we discuss energy consumption, security treats, CAPEX/OPEX, performance, spectrum and mobility management aspects of 5G C-RAN. Moreover, we review existing survey and white papers, which explore various dimensions of C-RAN. Finally, we summarize the review of state-of-the-art related to C-RAN in Table 3.

1) ENERGY CONSUMPTION

Due to centralization of baseband processing functions in C-RAN, the number of BS sites is reduced to several folds. Therefore, the number of air conditioning and site support equipment is also decreased [73], [75]. On the other hand, distance among RRH and UE is reduced thanks to cooperative radio technology that is also used to lower interference among RRHs. Both of these unique characteristics of C-RAN lead to decrease energy consumption in 5G mobile network.

Energy consumption in C-RAN has been studied to some extent, however, it is still in nascent stage that needs further research. In [76], the authors discover the minimization of downlink power in C-RAN considering coordinated transmission problem. Whereas [77] proposes a resource utilization framework that is claimed to dynamically change transmit power in order to meet capacity demand of end-users. Efficient energy saving schemes are also crucial in C-RAN. Hence, the results obtained in [78] demonstrate that their proposed energy saving schemes save significant amount of energy in comparison to traditional RANs.

Reducing computation complexity in C-RAN plays a significant role in the decreasing of energy consumption. This has proved in [79], where the authors propose a pre-coding scheme aiming to reduce computation complexity. Energy consumption in the C-RAN is also analyzed based on queuing theory using Ethernet-based Time Division Multiplexing

Passive Optical Network (TDM-PON) as optical fronthaul in [80]. Moreover, power consumption of different RANs assuming various level of centralization of baseband functions is examined in [81]. In [82], [83] a comparative analysis of power consumption of various C-RAN deployment scenarios is provided. The results claim that C-RAN saves 75% of total energy consumption. Last but not least, [84]–[86] explore energy efficiency of fully-centralized C-RAN considering power consumption of radio and transport segments.

2) SECURITY

Number of cyber-attacks are increasing day by day, therefore network security is a concern for C-RAN as well. C-RAN architecture faces various security threats such as eavesdropping and jamming in the physical layer, using Media Access Control (MAC) spoofing to impact Network Interface Controller (NIC) of various nodes' MAC addresses in the MAC layer, IP spoofing and IP hijacking in network layer, Transmission Control Protocol (TCP) flooding attacks in the transport layer, and malware and File Transfer Protocol (FTP) attacks in the application layer [87].

So far, a small number of studies has been carried out to explore possible solutions to security threat in C-RAN [75], [87]–[90]. In [87], the authors present a detailed survey on C-RAN security with primary focus on security threats and attacks based on three logical layers namely, physical plane, control plane, and service plane that are proposed as parts of logical structure of C-RAN in [75]. Reference [90] proposes wireless channel security with channel estimation errors. Whereas, [89] investigates security threats related to spectrum. Lastly, the authors of [88] propose a user-centric security resource allocation mechanism in the C-RAN.

3) CAPEX/OPEX

The splitting of L1, L2 and L3 functions among BBU and RRH reduces CAPEX and OPEX in C-RAN. There exists a considerable body of literature on both expenditure and operational costs of C-RAN. For example, the authors in [69] analyze that 80% of total CAPEX of a network operator is spent on RAN architecture and 41% of total OPEX of a single cell site is spent on power consumption. Moreover, the CMRI in [73] finds that by adapting of C-RAN, 15% of CAPEX and 50% of OPEX reduction can be achieved.

The deployment costs of BSs, transport networks, and data center in C-RAN are analyzed in [91] using various spatial point processes. In [92], the authors propose a framework in order to find out optical BS clustering scheme in C-RAN. The results show that proposed framework reduces 12.88% of deployment cost. Furthermore, [93] models and compares the cost of three BBU strategies including BBU stacking, BBU pooling, and C-RAN BBU. The results obtained in the paper show that C-RAN BBU is appropriate solution from cost saving perspective. Lastly, the authors in [94] propose a network planning framework, which optimizes the cost deployment of C-RAN in 5G mobile networks.

4) PERFORMANCE

The deployment of C-RAN has attracted significant attention due to its high flexibility and performance. The transmission link between BBU and RRH in C-RAN is either dedicated fiber or microwave, therefore, high bandwidth services can be delivered in ultra low latency, which leads to increase overall performance of the system [95]. Moreover, with the deployment of C-RAN, the downlink data rates at the cell edge can be improved up to 40%-70%, and uplink data rates at the cell edge can be improved by up to 2-3 folds [95].

The performance of various 5G RANs including C-RAN is evaluated in [96], whereas, [90] investigates reliability performance of downlink in the C-RAN considering Channel Estimation (CE) errors. Furthermore, in [97], the authors investigate performance improvement of C-RAN aiming to minimize power consumption of the network. However, [98] analyzes the performance of C-RAN, where RRHs are assumed to aid macro BS in the transmission.

5) SPECTRUM

The spectrum is approaching to its theoretic limit, on the other hand, deployment of various technologies in the RAN architecture of 5G system makes the network dense [75]. Therefore, operators are forced to find efficient mechanisms in order to increase spectrum efficiency in the future networks. As discussed earlier, one of the possible ways to increase spectrum efficiency in 5G mobile network is the deployment of C-RAN. So far, a few number of investigations have been carried out by many researchers aiming to increase spectrum efficiency in the C-RAN. Among them, [99] investigates the pros and cons of licensed and unlicensed spectrum of various candidate technologies for the fronthaul of C-RAN towards 5G mobile network.

6) MOBILITY MANAGEMENT

Among numerous advantages of C-RAN, enhanced mobility management is considered one of the key topics that has extensively been studied. In [100], the authors propose location based algorithms in order to cluster towers, and pack BBU clusters based on mobility and traffic patterns prediction in the C-RAN. The results in the paper claim that 34.8% improvement in QoS is obtained. Reference [101] explores that dynamic assignment of computing resources in the C-RAN generates a new class of handover, which has negative impact on the QoS. The authors claim that based on the proposed algorithm, the total number of handovers in C-RAN is reduced by 20% in comparison to traditional RAN architecture. On top of that, the authors in [102] analyze handover performance in C-RAN. The results in the paper claim that C-RAN architecture plays a significant role in decreasing of handover delay, moreover, it eliminates the risk of end-user losing its connection.

7) SURVEY AND WHITE PAPERS

Despite above discussed technical papers, a number of surveys related to the C-RAN has been published so far.

TABLE 4. A summary of contributions by mobile operators and vendors to the C-RAN in 5G networks.

Reference	Organization	Contributions to the C-RAN
[73]	China Mobile	Discusses concept, benefits, challenges, architecture, technology trends, virtual BS, fully/partially centralized C-RAN, and dynamic resource allocation.
[74]	IBM	Explores system requirements, architecture, application scenarios, virtualization of BS pool, and open research challenges.
[95]	Telefonica	Addresses virtualization of RAN, NFV/SDN in RAN, centralized and distributed RANs,splitting of functions of RAN in 5G, and transport networks.
[104]	Fujitsu	Network coverage, capacity expansion, research challenges, transport options, implementation, and architecture are discussed.
[105]	Ericson	Legacy RAN architectures, distributed baseband, centralized baseband, architecture, virtualization, centralization, and coordination in the C-RAN are explored.
[106]	Huawei	Deployment, advantages, different RATs, system architecture, deployment of C-RAN in 5G, and integration of Edge-Cloud in the C-RAN are investigated.
[107]	ZTE	Virtualization platform based C-RAN architecture and C-RAN characteristics.
[108]	NEC	Architecture, functions splitting, and cell virtualization.
[109]	Intel	Advantages and challenges, Intel C-RAN solution, and virtualization of 5G cells.
[110]	Texas Ins.	The position of cloud BS in the RAN architecture, system interconnection, SDN and NFV in the RAN, virtualization technology and standards.

These survey papers explore C-RAN from various perspectives such as security [87], architecture [69], [75] and NFV/SDN [103]. On the other hand, the world well-known MNOs and vendors including China Mobile [73], IBM [74], Telefonica [95], Fujitsu [104], Ericson [105], Huawei [106], ZTE [107], NEC [108], Intel [109], and Texas Instruments [110] have also investigated various dimensions of C-RAN, we have summarized their key contributions in Table 4.

To conclude, mobile operators can benefit from various advantages of C-RAN such as reduced CAPEX/OPEX, decreased power consumption, low latency, improved efficiency in resource allocation, increased flexibility during network upgrading, and enhanced adaptability to non-uniform traffic. Moreover, there are many other key advantages that C-RAN brings to 5G wireless communication system. All these advantages are thoroughly discussed in [72], [73], [104]. Despite above-mentioned key advantages, the deployment of C-RAN in 5G mobile networks leads to various limitations, which open new research challenges that are needed to be addressed before its implementation. Some of the major limitations include network security, transportation latency, cooperative transmission and reception, virtualization of BS, etc. We provide a detailed discussion on some of the major research challenges related to the C-RAN in Sec. X.

V. HETEROGENEOUS CLOUD RADIO ACCESS NETWORK

A. HETEROGENEOUS NETWORK

The RAN architecture of 5G mobile network is more heterogeneous in comparison to the RAN architectures of legacy LTE/LTE-A networks. The authors in [60] predict that density of BSs in 5G RAN is highly anticipated to come up to $40-50$ BSs/Km^2. Moreover, data traffic demand in cellular networks is also increasing day by day. Therefore, further improvements in system capacity and spectral efficiency are needed to fulfill the requirements of end-users beyond 2020. One possible way to achieve this goal is to deploy small cells over existing macrocellular layout, which is called Heterogeneous Cellular Layout. The well-known theory of Shanon shown in (1) does also express this fact. As shown, approximate total capacity of a system is (C_{sum}). The B_i is the bandwidth of ith channel, Pi denotes the signal power of the ith channel, and Np is the noise power. It is clear that C_{sum} is equivalent to the sum capacity of all subchannels. In order to increase C_{sum}, we need to deploy more macrocells and small cells, which leads to a heterogeneous cellular environment (see Fig. 14). In a heterogeneous cellular layout, the macrocells tier is used to provide wide coverage and seamless mobility over large geographic areas, whereas small cells are deployed to improve coverage and increase capacity by moving computation and communication nodes close to the end-users.

$$C_{sum} \approx \sum_{HetNets} \sum_{Channels} B_i Log_2 \left(1 + \frac{P_i}{N_p}\right) \tag{1}$$

The small cells/low-power nodes encompass a broad variety of cell types, such as femtocell, picocell, and microcell [111]. These low-power cells are deployed in various environments, i.e., hot spots, homes, enterprise environments, shopping malls, stadiums, train stations, and other

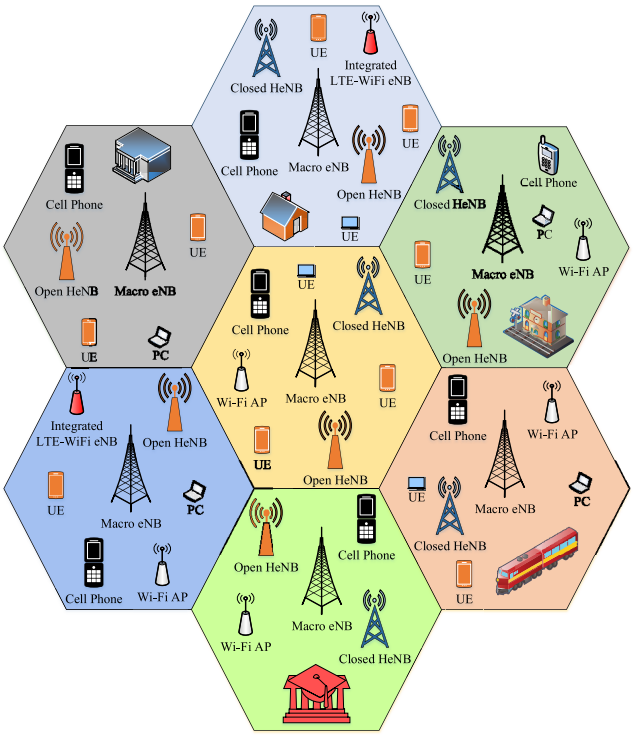

FIGURE 14. Heterogeneous cellular layout.

small geometrical areas in order to improve overall network capacity and coverage, decrease network cost, and furthermore increase spectral efficiency. As shown in Fig. 14, various types of small cells are aggressively deployed to address capacity and coverage demands of future 5G networks.

Small cells can be deployed both indoor and outdoor. Those small cells, which are intended to be deployed outdoor require low power that ranges from 250 mW to approximately 2 W. Small cells are available in lower cost in comparison to traditional macro BSs, which require high power range from 5 W to 40 W [112]. Those small cells, which are intended to be deployed indoor require transmit power of typically 100 mW or less. The coverage area of small cell is significantly smaller in comparison to macrocells due to their lower transmit power, which can limit the volume of data [113].

Indoor femtocells may be configured with/without restrictions. The femtocell, which is strictly configured and allows only specific users to access is called Closed Femtocell (Closed HeNB). The femtocell, which can be accessed by any of the users in a specific geographic area is called Open Femtocell (Open HeNB). There is also a Hybrid Femtocell, which allows unsubscribed users to access with an upper limit on the amount of the available resources. The Closed HeNB configures resection list through Closed Subscriber Group (CSG) [112].

Small cells and heterogeneous cellular layout have been widely studied in the literature. Recently, an outstanding source for understanding of small cells in 5G mobile network has been published by Cambridge University Press [114]. The book covers all aspects of small cells network including designing, optimization, and deployment. It details fundamental concepts and advanced topics related to the small cells. The reference furthermore covers emerging trends, research challenges, resource management, energy efficiency, performance analyzing, deployment strategies, standardization activities, environmental concerns, and mobility management in small cell heterogeneous networks. On the other hand, [115] details the creation and deployment of small cells, and the technical challenges associated with the design, deployment and optimization of small cell network. The authors discuss critical technical elements such as coverage and capacity optimization, mobility management, interference management, energy efficiency, backhaul, deployment planning, frequency assignment and access methods, and heterogeneous networks management.

There have been numerous amount of studies conducted on femtocells deployment. Most of these studies are dealing with operations, administration, management, access, interference management, local IP access (LIPA) and architecture that can be found in [116]–[119]. Low OPEX/CAPEX is one of the main driving forces of deployment of femtocell. As shown in [120], in dense urban area, a combination of open HeNB (randomly deployed by end-users) and macro eNB (deployed by an operator in a planned manner) can result up to 70% of the total annual network cost in comparison to a pure macrocellular layout. The performance evaluation of heterogeneous network composed of macro eNB, pico eNB, and open HeNB is evaluated in [121], [122]. The results obtained in both references show that the main reason behind limited performance gain of heterogeneous network is the limited coverage of low power nodes.

B. H-CRAN ARCHITECTURE

Recently, a new RAN architecture called the Heterogeneous-CRAN or H-CRAN has been proposed to decouple both control and user planes in order to enhance the functionalities and performance of C-RAN architecture, in which control plane functions are only implemented in the macro BSs. In H-CRAN, full advantages of both heterogeneous network and C-RAN have been taken, which lead to improve spectral and energy efficiencies and meanwhile enhance the data rates. The H-CRAN architecture consists of two cellular layouts [123], [124], the macro BSs (high power nodes) cellular layout and the small BSs or RRHs cellular layout. The High Power Nodes (HPNs) are mainly deployed to enhance network coverage and furthermore control network signaling. However, the small cells and RRHs are aimed to guarantee improved network capacity and fulfill the diverse requirements of QoS of various end-users. As depicted in Fig. 15, the system architecture of H-CRAN is consisted of three main functional modules [124]:

- **Enhanced-Cloud and Real-Time Virtualized BBU Pool:** In this functional module, all BBUs scattered to different cells are integrated into the BBU pool.

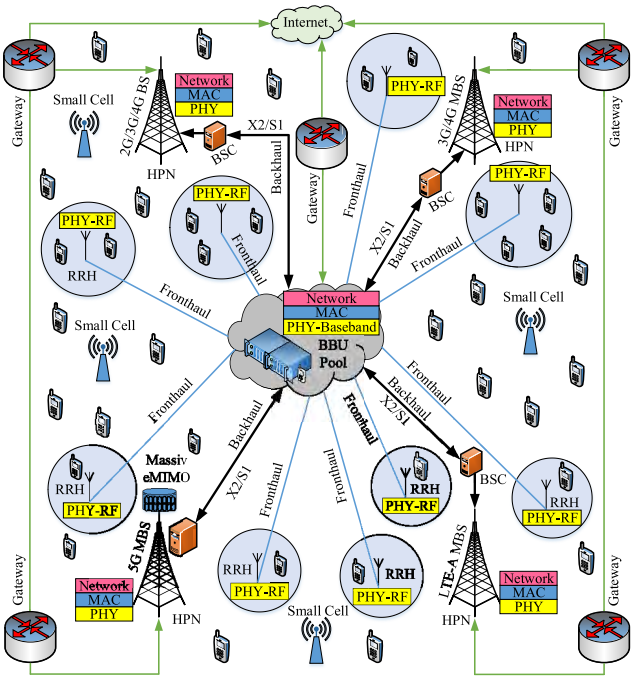

FIGURE 15. System architecture of H-CRAN.

The BBU pool is built up on strong cloud-computing and powerful virtualization techniques. Moreover, the BBU pool is connected to HPNs in order to coordinate the inter-tier interfaces between HPNs and the RRHs.

- **Extremely Reliable Transport Network:** As shown in Fig. 15, all RRHs are connected to its corresponding BBUs located in the BBU pool. Both RRHs and BBUs are interconnected via low latency and high bandwidth fronthaul links such as optical fiber. The S1 and X2 are the data and control interfaces between BBU and MBSs (HPNs), respectively.
- **High Number of Macro BSs, Small BSs and RRHs:** In H-CRAN architecture, different types of cells including macro base stations, small base stations, and RRHs are coexisted. The macro base stations control network, manage mobility, and improve performance; whereas small cells and RRHs increase system capacity and decrease transmission power. The symbol processing and radio frequency functionalities are configured in the RRHs, however, all other baseband physical processing functionalities of upper layers are integrated into the BBU pool. The high power nodes are equipped with all functionalities from the physical layer to the network layer.

In H-CRAN, the enhanced cloud computing centralized integration of all BBUs, the separation of functions between RRHs and BBUs, and the decoupling of control plane and the data plane lead to efficient management of heterogeneous mobile networks. Therefore, in scenarios such as expanding network coverage and improving system capacity, the mobile operators need to only install new RRHs close to the user and furthermore connect them to the BBU pool. Moreover, the implementation of flexible software solutions is also fairly easy. For example, if network operator is interested to upgrade RANs and support multi-standard operations, then, it can be possible only through updating softwares by deploying SDR.

From the architectural point of view and as depicted in Fig. 15, there are two main similarities between traditional C-RAN and H-CRAN: (a) for the sake of achieving high cooperative gain and increasing energy efficiency, a huge number of RRHs are connected to a centralized BBU pool. (b) RF and simple symbol processing functionalities are executed in the RRHs, however, all those functionalities that are related to upper layers are implemented in the BBU pool. Despite two above mentioned common similarities between traditional C-RAN and H-CRAN, there are some differences that are needed to be addressed including: (a) as illustrated in Fig. 15, the BBU pool and the HPNs are interfaced (denoted as S1 and X2 respectively) in the H-CRAN in order to mitigate the cross-tier interference between RRHs and HPNs coexisted in the same geographical area. (b) With the participation of HPNs, the requirements of the fronthaul are decreased in the H-CRAN such as system broadcasting data and the control signaling are delivered to the UEs through HPNs, which leads to simplify capacity and time delay in fronthaul links. (c) When the traffic load of an RRH is low in the H-CRAN, the RRH falls into sleeping mode in order to improve energy efficiency, however, the BBU pool manages all those RRHs, which are in the sleeping mode.

C. STATE-OF-THE-ART

The H-CRAN has been researched from different perspectives such as energy consumption, spectrum efficiency, inter-tier interference, architecture, mobility management, performance, etc. We have summarized the contributions of published papers studying H-CRAN from different perspectives in Table 5. Moreover, a detailed description of their main contributions is provided in the following.

1) ENERGY CONSUMPTION

The energy efficiency of H-CRAN has been explored in [123]–[133]. In [125], the authors study a joint optimization solution for the assignment of resource blocks and allocation of power in order to maximize the performance of energy efficiency in OFDMA-based H-CRAN. While the same authors in [123] discuss system architecture, spectral and energy efficiency performance, and key promising techniques of H-CRAN. Paper [124] explores green evolution of H-CRAN from three different perspectives; energy efficiency-spectrum efficiency, energy efficiency fairness, and energy efficiency delay trade-offs. The authors in [126], [127] analyze and evaluate power consumption of C-RAN and H-CRAN in order to enhance energy and spectral efficiencies. Results obtained in both papers show that with the deployment of C-RAN, a network operator reduces about 87% of the cooling power consumption, which

TABLE 5. A summary of major works related to H-CRAN in 5G mobile networks.

Research Issues	Briefing	References
Energy Efficiency	Discuss power allocation to nodes and transport networks, and key promising techniques in order to maximize the efficiency of energy consumption in the H-CRAN.	[123]–[133]
Performance	Explore the data rate on both downlink and uplink, study mechanisms to fulfill the QoS requirements of different users, and propose schemes to enhance throughput.	[123], [125], [129], [132] [134]–[141]
Resource Allocation	Study efficient allocation of network resources and assignment of sub-channels.	[128], [130], [133], [139] [140], [142]–[144]
System Architecture	Investigate the architecture and topology along with key enabling technologies.	[124], [125], [142], [145] [146]
Spectrum Efficiency	Address the efficiency of spectrum to improve the performance.	[123], [124], [126], [127]
Interference	Propose solutions to mitigate inter-tier interference.	[135], [147], [148]
Surveys	Survey available contributions and key technologies.	[145]
CAPEX/OPEX	Develop framework to decrease both CAPEX and OPEX.	[91]
Load Balance	Balance traffic load on different RRHs and transport links.	[149]
Mobility	Propose cell selection and mobile equipment mobility features.	[150]

decreases the overall power consumption by about 40% compared to macro BS deployment. In [128], the authors propose a resource-optimal scheme in order to achieve low latency guarantees for real-time applications using minimum amount of replicated radio resources. The paper suggests a system architecture in order to increase throughput of non-real-time applications.

Resource optimization and joint congestion control are crucial issues in H-CRAN, thus, the authors in [129] study both of these aspects of H-CRAN in order to investigate energy efficiency-guaranteed trade-off between delay performance and throughput utility in a downlink of H-CRAN using stochastic optimization problem. The authors in [130] discuss an energy-efficient optimization objective function with individual fronthaul capacity and intertier interference constraints for queue-aware multimedia H-CRAN. In [131], the authors investigate transmit power of RRHs and HPNs in order to maximize energy efficiency in the H-CRAN. Moreover, the authors have developed the optimization problem with total allowed power, and the optimal power allocation for each of the RRHs and HPNs are derived through dual decomposition method. In paper [132], the authors study the downlink problem between cloud-base central station and multiple BSs in the context of H-CRAN using the same time and frequency resources. The authors analyze energy efficiency utilizing NOMA and considering channel modeling with power consumptions at different type of cells in the context of H-CRAN. Last but not least, an algorithm that optimizes energy efficiency of radio resource allocation for H-CRAN is proposed in [133].

2) PERFORMANCE

The performance of H-CRAN has been widely studied in the literature. Most of the papers published so far in this area are listed in Table 5 including [123], [125], [129], [132], [134]–[141]. As previously discussed, [123], [125], [129], and [132] evaluate the performance of H-CRAN from energy consumption perspective. However, in paper [134], the authors study the problem of cooperative radio resource management in the H-CRAN emphasizing on its real-time performance optimization. Paper [135] considers inter-tier interference in order to analyze ergodic rates of downlink in the H-CRAN for two access methods proposed by the authors - the distance based and the cluster based. The authors in [136] propose non-uniformly D2D communication integrated with H-CRAN, where the D2D links are utilized outside a specified distance from any HPN in order to offload RRHs located in the coverage area of HPNs. In [137], the authors study a mechanism to enhance throughput of a single cell in H-CRAN considering cooperative communication among HPN and LPNs. The paper specifically focuses to form a cooperative cluster among LPNs for cross-tier cooperation, and has furthermore proposes a hierarchical cooperation strategy in order to improve throughput. Paper [138] investigates the provisioning of QoS-guaranteed services to as many users as possible in the H-CRAN. The main objective

of the paper is to optimize the number of end-users, each user with its own data rate requirements considering a given number of BSs with specific bandwidth and transmission power.

Efficient resource allocation has significant impact on the enhancing of performance in the H-CRAN. In [139], the authors propose a resource allocation scheme in two approaches - centralized and decentralized, based on online learning, which aims to mitigate interference, maximize energy efficiency, and meanwhile maintain QoS requirements of various end-users in the H-CRAN. In [140], the authors propose a centralized resource allocation scheme using online learning to guarantee interference mitigation, increase energy efficiency, and fulfill the QoS requirements of all types of users in the H-CRAN. Last but not least, in paper [141], the authors propose a radio resource management scheme aiming to optimize the energy efficiency of the H-CRAN. They have developed an energy consumption model, which is used to characterize energy consumption of RRHs, fronhaul links, and the BBU pool in the H-CRAN.

3) RESOURCE ALLOCATION

Resource allocation of H-CRAN has also been widely studied by researchers in [128], [130], [133], [139], [140], [142]–[144]. We have briefly described the contributions of [128], [130], [133], [139], [140] previously. However, the authors in [142] explore existing research challenges and recent developments related to the design of H-CRAN. The article proposes resource allocation schemes including coordinated scheduling, hybrid backhauling, and multi-cloud association. The authors of [143] study sub-channels assignment of different bandwidth to various D2D pairs and the users of RRH of H-CRAN of 5G mobile network. By using this scheme, all pre-allocated macro-cell sub-channels could be used, which leads to enhance system performance and guarantee the QoS for all users. In [144], the authors explore resource sharing in H-CRAN at three levels namely the spectrum, infrastructure and network. The article further shed light on the advantages/disadvantages and promising technologies including SDN, NFV, and SDR, which make resource sharing possible and feasible in the H-CRAN.

4) SYSTEM ARCHITECTURE

The system architecture of H-CRAN has been thoroughly studied in [124], [125], [142], [145], [146]. We have explored the contributions of [124], [125], [142] in previous paragraphs. However, in [145], the authors survey state-of-the-art contributions related to key technologies and system architecture of the H-CRAN. Meanwhile, in [146], the authors analytically derive the optimum collaborative access for both H-CRAN and EdgeNet architectures.

5) SPECTRUM EFFICIENCY

The spectrum efficiency is considered one of the important topics of H-CRAN that has attracted signification attention from the research community. Those papers, which address the efficiency of spectrum in the H-CRAN are available in [123], [124], [126], [127], which have been thoroughly described in the previous paragraphs.

6) INTERFERENCE

The authors in [135], [147], [148] study interference in the H-CRAN. We have explored reference [135] previously, where the authors analyze ergodic rates on downlink considering inter-tier interference. However, in [147], the authors propose a contract-based interference coordination framework in order to mitigate the inter-tier interference between RRHs and HPNs in the H-CRAN. The authors in [148] address the inter-tier interference techniques considering collaborative processing and Cooperative Radio Resource Allocation (CRRA). Moreover, Beamforming and interference collaboration are proposed to suppress the inter-tier interference between RRH and HPN.

7) CAPEX/OPEX, LOAD BALANCING AND MOBILITY

Despite above mentioned aspects of H-CRAN, there are some papers, which discover CAPEX/OPEX, load balancing, and mobility management including [91], [149], [150]. In [91], the authors propose a theoretic framework that enables computation of the deployment cost of a H-CRAN. The results in the paper demonstrate that cloud-based RANs require approximately 10% to 15% less CAPEX per square kilometer in comparison to traditional 4G networks. Article [149] researches link capacity between RRHs and the BBU pool (backhaul) in order to balance traffic load, enhance backhauling performance, and decrease pressure on the transport links. As for the mobility management and more specifically the cell selection, the authors in [150] propose a scheme for the user equipments to directly select the cell with the highest priority in high dense H-CRAN. The priorities, which are considered in the paper are cell features, mobile equipment mobility features and application features.

So far, there have been many progress and achievements in terms of performance analysis, system architecture, key techniques, potential applications, and various other aspects of H-CRAN. However, there are still many research challenges and open issues including optimal resource allocation, standards development, theoretical performance analysis, balancing backhaul load, spectrum/energy efficiencies, and so on. We provide a detailed description of major research challenges related to the H-CRAN in Sec. X.

VI. VIRTUALIZED CLOUD RADIO ACCESS NETWORK

NFV has drawn significant attention from research community due to its remarkable advantages in increasing the efficiency of network resource sharing and enhancing the flexibility of scheduling. The virtualization in telecommunication networks has been deployed for the first time in core network and later on extended to the radio access domain, which is still in its nascent stage. However, the distinctive characteristics of wireless communication system such as mobility, broadcast, attenuation, interference, time-various

channels, coverage, etc., make the virtualization of wireless network more challenging.

The virtualization of cellular network includes various scopes of virtualization including infrastructure virtualization, spectrum virtualization, air-interface virtualization, virtualization of multiple radio access technologies, and virtualization of computing resources. Virtualization in cellular network enhances the efficiency of resource utilization, improves network performance, simplifies the migration to new technologies, reduces CAPEX/OPEX, increases the revenue, and creates new businesses in the market.

Despite all benefits that virtualization brings to the wireless communication industry, it also faces three main challenges that are needed to be addressed: a) Wireless resources must be shared and distributed to different virtual network operators in an efficient and fairly manner. b) The interferences that are caused due to the resource utilization should be strictly considered. c) Before deployment of virtualization in the wireless network, many technical and managerial related problems should be investigated.

In order to enable end-to-end virtualization in the 5G cellular network, both the CN and RAN architecture have to be virtualized. The virtualization solution for the wired network can also be deployed in the virtualization of the CN of a mobile network. However, the virtualization of RAN or BS has to be performed considering specific characteristics related to the wireless access network. The dynamic sets of end-users, mobility, and varying channel conditions require new virtualization mechanisms than those of wired ones. Therefore, these characteristics make the virtualization process harder for mobile operator to virtualize wireless network resources.

With the deployment of C-RAN, network related resources are deployed close to the end-user, making it easier to implement all core network related functions and applications in a cloud data center. Mobile operators are linking data center to the MEC in order to lower the latency and improve the performance. However, massive number of users' devices, lower latency services, and higher capacity demands of users and vertical industries are the requirements that 5G mobile network is expected to fulfill. A good response to all these requirements would be to deploy NFV and SDN in the C-RAN in order to virtualize all functions and resources in the RAN architecture and furthermore decouple data and control planes. This new trend of virtualization towards access network forms a new type of C-RAN, which is called Virtualized-CRAN or V-CRAN.

A. SYSTEM ARCHITECTURE OF V-CRAN

In Fig. 16, an example of physical architecture of V-CRAN is provided, which is composed of various subsections such as Digital Unit Cloud (DU Cloud), Time-Wavelength-Division Multiplexed Passive Optical Network (TWDM-PON), fronthaul, and Virtualized Base Station (V-BS) [151]. In the DU Cloud, all commercial servers are deployed, which provide baseband processing, layer 2 and layer 3 related functions. Moreover, all DUs are interconnected with each other through

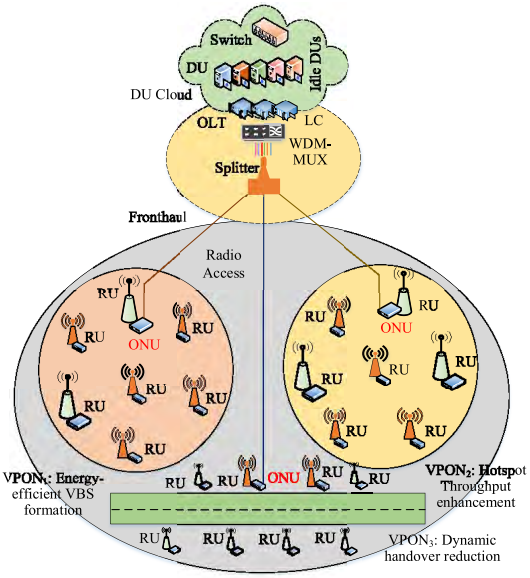

FIGURE 16. System architecture of V-CRAN.

a layer 2 switch that is used to exchange data and signaling among the DUs.

The V-CRAN requires higher bandwidth and lower latency fronthaul network in order to exchange data and signaling between Radio Units (RUs) and DUs. To fulfill all these requirements, utilization of optical channels over single fiber is considered one of the viable candidates of V-CRAN's fronthaul. The fronthaul shares high-capacity optical channels among V-BSs using TWDM-PON technology. The TWDM-PON is composed of a single Optical Line Terminal (OLT) and multiple Optical Network Units (ONUs). As illustrated in Fig. 16, the OLT is connected with the DU cloud. It is responsible to provide each DU an optical transceiver and a Line Card (LC), which conducts optical-electrical conversion and furthermore deliver the traffic. All LCs of every single OLT are connected with a Wavelength-Division Multiplexed Multiplexer/De-Multiplexer (WDM-MUX) in order to separate traffic based on wavelength. An ONU is located remotely from DU Cloud and at the end of every optical channel aiming to enlarge the coverage of a TWDM-PON. It is co-located with an RU and is equipped with a reconfigurable transceiver.

In the access network, ONUs share one or more wavelengths in order to form a Virtualized Passive Optical Network (VPON). The VPON is a virtualized channel that is used to connect RUs with the DU. Every DU in the DU Cloud is dedicated to a single VPON using LC. It specifically means that multiple RUs of a VPON are dynamically grouped and controlled by a single DU located in the DU Cloud. When a V-BS is highly loaded during the busy hours, each cell is assigned with a dedicated VPON. However, during the idle hours, the cells are grouped together to form a VPON. Therefore, OPEX is decreased, energy is consumed less, and network resources are utilized efficiently.

TABLE 6. A Summary of major works related to V-CRAN in 5G mobile networks.

Research Issues	Briefing	References
Architecture	Investigate system architecture and topology. Explore key enabling technologies along with recent advancement on SDN and NFV and its implication in V-CRAN.	[103], [151], [153], [154], [155]-[158]
Performance	Discover throughput on downlink and uplink, study mechanisms to fulfill QoS requirements of different users, and propose schemes aiming to enhance throughput.	[151], [152], [159]-[163]
Resource Allocation	Investigate efficient allocation of network resources and its assignment to different nodes of the network.	[151], [154], [158], [163]-[165]
Surveys	Survey available contributions and literature related to V-CRAN.	[103], [151], [152], [167]
CAPEX/OPEX	Develop framework to decrease both CAPEX and OPEX.	[152], [166], [167]
Energy	Explore power allocation techniques aiming to maximize energy.	[165], [166]

Recently, a new concept of V-BS was proposed in order to virtualize computing resources of a BS in V-CRAN [151]. The virtualization of a BS is performed at two distinct levels: a) hardware level (dedicated spectrum). b) flow level (shared spectrum) [152]. The V-BS shares radio equipment at hardware level and runs multiple protocol stacks of a BS in the form of software. The hardware virtualization solution has already been standardized and traditional mobile operators have been using this scheme to decrease OPEX and increase energy efficiency. As for the spectrum sharing-based models, virtualization at higher levels such as flow level at the V-BS is needed in order to increase the efficiency of resource multiplexing. Moreover, the spectrum sharing-base models support various scenarios such as the deployment of mobile virtual network operators which does not own the spectrum.

B. STATE-OF-THE-ART

V-CRAN has attracted significant attention from both academia and industry. There has been numerous amount of studies, which explore V-CRAN from different aspects such as energy consumption, resource allocation, CAPEX/OPEX, performance, and system architecture. In the rest part of this section, we have tried to the best of our knowledge to review up to date literature related to all these dimensions of V-CRAN. A review of this literature is summarized in Table 6 and its detailed descriptions are provided in the following.

1) SYSTEM ARCHITECTURE

The system architecture of V-CRAN has been thoroughly explored in [103], [151], [153], [154], [155]–[158]. Reference [103] surveys virtualization technology in next generation of wireless networks considering its recent advancement in the SDN and C-RAN. The authors further propose a general architectural framework for the virtualization of wireless network based on SDN. The authors of [151] present the physical architecture of V-CRAN and the concept of V-BS for 5G mobile network. In [153], the authors propose V-BS architecture considering OFDM air interface and furthermore investigate the flexibility of resource allocation in the radio layer. Paper [154] thoroughly describes the requirements, physical architecture, and concept of V-BS in V-CRAN. The authors further demonstrate a cloud management solution for the V-CRAN. The authors in [155] propose an architecture for V-BS in the V-CRAN. On the other hand, papers [156], [157] propose a Not Only Stack (NO Stack) based V-CRAN candidate for 5G mobile network. The authors analyze various use-cases in order to validate the advantages of the proposed architecture. In paper [158], the authors design a wireless network virtualization architecture that is composed of three distinct planes; namely the cognitive plane, the control plane, and the data plane aiming to optimize dynamic allocation of network resources. They have further implemented a trail environment of proposed architecture claiming that it increases spectrum efficiency two times and reduces packet loss to 1/20.

2) PERFORMANCE

The system level performance of V-CRAN is analyzed in [152], [159]–[162]. In [152], the authors review state of the art research on virtualization of mobile carrier focusing on RAN sharing, and the programmability and virtualization of BS. Moreover, the authors present a RAN sharing technique and evaluate its performance in the context of LTE network. In paper [159], the authors present a virtualization framework of BS in the LTE network called "OpeNB". The proposed framework utilizes SDN, OpenFlow, and virtualization technology in the RAN and more specifically in the BS. In [160], the authors propose a framework, which assigns limited rule space in order to maximize the number of social IoT groups and furthermore satisfy latency in the context of V-CRAN. Paper [161] analyzes the performance of V-CRAN in terms of throughput, system stability and fairness considering two resource allocation strategies: the

Dominant Resource Fairness (DRF) and Proportional Fairness (PF). In [162], the authors design two types of performance metrics namely the macro level metrics and the micro level metrics of virtualization of BS. Moreover, the paper proposes an evaluation method for the BS virtualization platforms. Meanwhile, [151], [163] partially investigate system level performance of V-CRAN together with system architecture.

3) RESOURCE ALLOCATION

Resource allocation of V-CRAN is one of the major aspects that has widely been studied in [151], [154], [158], [163]–[165]. Reference [163] proposes an algorithm of resource negotiation for network virtualization. The algorithm is applicable in the context of heterogeneous LTE-A network aiming to achieve slicing and on-demand delivery of radio resources. The authors in [164] investigate air interface resource virtualization, coordination, and its dynamically allocation by a hypervisor among different virtual operators in V-CRAN. Reference [165] provides a resource allocation framework for V-CRAN and furthermore explores energy consumption. Moreover, [151], [154], [158] do also partially explore resource allocation together with system architecture and system level performance of V-CRAN, respectively.

4) ENERGY CONSUMPTION

The energy consumption of V-CRAN has been studied in [165], [166]. The authors of [165] propose a dynamic resource allocation framework namely KORA for heterogeneous V-CRAN architecture of 5G mobile network. The simulation results demonstrate that the heuristic algorithm is able to save 27.39% of number of relocations and 33.26% of number of affected Guaranteed Bit Rate (GBR) flows, and increases the consumption of energy by 6.67% in comparison to KORA. In paper [166] a cost effective scheme for V-CRAN is proposed. The authors claim that proposed scheme reduces average power consumption by 65%, 6% and 3% less than for the Distributed Baseline (DB), the First Fit Decreasing (FFD) algorithm and the Heuristic Simulating Annealing (HSA) algorithm, respectively.

5) CAPEX/OPEX

The CAPEX/OPEX of V-CRAN has been studied in [152], [166], [167]. The authors in [167] explore advantages, existing challenges, limitations and standardization of deployment of virtualization in V-CRAN for 5G mobile network. They analyze the cost of these requirements in V-CRAN. Moreover, [152], [166] do also partially analyze the CAPEX/OPEX together with system level performance and energy consumption, respectively. As for the survey and tutorials related to V-CRAN, we recommend interested readers to [103], [151], [152], [167].

To sum up, MNOs can take full advantages of various benefits of V-CRAN such as decreased power consumption, reduced CAPEX/OPEX, isolation of different services each with its own virtual network, flexibility during network upgrading, and adaptability to non-uniform traffic. Despite all these advantages that V-CRAN brings to 5G systems, there are some challenges that come from its implementation. Most of these major research challenges are related to C-RAN, NFV, and SDN that effect on the deployment of V-CRAN. We discuss all these major research challenges in Sec. X.

VII. FOG-RADIO ACCESS NETWORK

The volume, variety and velocity of data that IoT and future networks generate are expected to increase at an unpretending rate. The International Data Corporation (IDC) predicts that there will be 30 billion sensor-enabled objects connected to the Internet, 110 million connected cars with 5.5 billion sensors, and 1.2 million connected homes with 200 million sensors by 2020 [168]. Cloud computing is one of the appropriate solutions, where remote servers hosted on the Internet, store, process, manage, and analyze the data generated by the IoT devices. It enables companies to focus on their core businesses instead of expending resources on planning, deployment, and maintenance of network infrastructure [169].

Current cloud computing techniques lead to many limitations such as large end-to-end delay, traffic congestion, processing of massive amount of data, and communication cost. Therefore, a new computing model for storing, processing, managing, analyzing, and acting on network data was proposed, which is called Fog Computing [170]. The term of ''Fog Computing'' was initiated by Cisco [171], which means ''fog is a cloud close to ground'' and it is used to extend the cloud computing to the edge of the network. In fog computing, computation, communication, control, and decision making processes are selectively moving towards the edge of the network in order to act on data in milliseconds, analyze time sensitive data close to the IoT device, and send selected data to the data center in cloud for longer-term storage or historical analysis.

Fog computing allocates a large amount of processing, storage, communication, control, configuration, measurement, and management functions at the edge of mobile network [172]. In fog computing, the Collaboration Radio Signal Processing (CRSP) can be executed in BBU pool and also hosted at the RRH and even further at the UEs (such as wearable smart UEs). In order to efficiently support and integrate new types of UEs, it is required that the on-device processing and Cooperative Radio Resource Management (CRRM) with less distributed storing should be exploited. However, from mobile applications perspectives, the UEs do not necessarily have to be connected to the BBU pool in order to download the packets if the they are locally available and stored in the closest RRHs.

Considering all aforementioned advantages of the Fog computing and alleviating the existing challenges of the H-CRAN, a new RAN architecture based on fog computing called the F-RAN was proposed, which splits a computation task into the fog computing part and the cloud computing part [173]. In this section, we focus on the physical architecture and literature review of the F-RAN. Therefore,

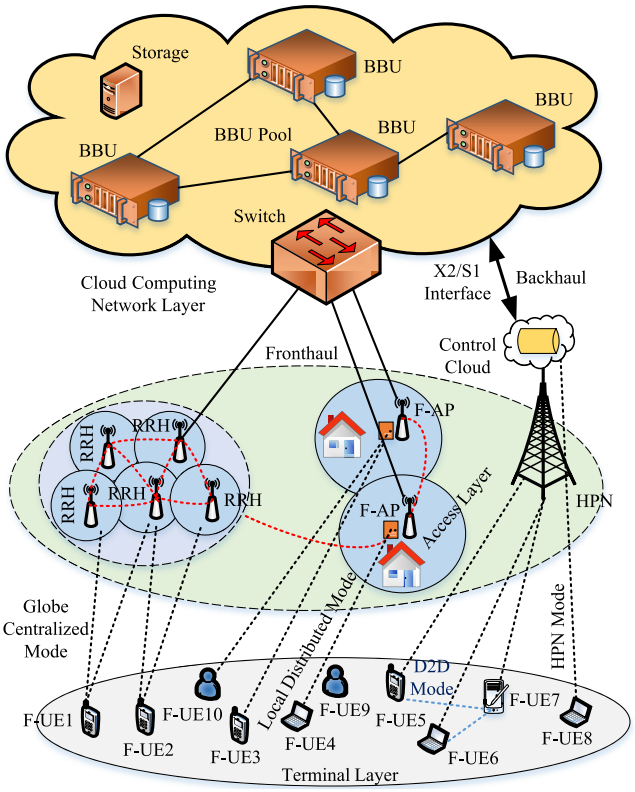

FIGURE 17. System architecture of F-RAN.

we skip detailed discussion on the applications, advantages and limitations, and system architecture of fog computing. However, if readers are interested to explore about various aspects of fog computing and its comparisons with cloud computing, then we refer them to [169]–[172].

The F-RAN was proposed aiming to take the advantages of both fog computing and C-RAN in order to tackle the dramatically increasing traffic demands and to provide better QoS to the end-users. The F-RAN is classified into two types [173], the Distributed F-RAN and the Centralized F-RAN. In distributed F-RAN, the BBU drifts some of its functions such as computation, resource management, and storage to the RRHs and the equipment of the users. Whereas, in the centralized F-RAN, the SDN and NFV are utilized to facilitate logically centralized control plan, and easier management and resource allocation.

A. SYSTEM ARCHITECTURE

A practical example of system architecture of F-RAN is depicted in Fig. 17, which is consisted of terminal layer, network access layer and the cloud computing layer [174].

- As shown, the Fog Access Points (F-APs) in network access layer and Fog User Equipments (F-UEs) in terminal layer formulate the mobile fog computing layer. In the terminal layer, the F-UEs do also access the HPN in order to receive information related to system signaling. Moreover, the neighboring F-UEs can communicate with each other in terminal layer using D2D communication mode. An example of this F-UE-based relay mode is shown in the terminal layer of Fig. 17, where F-UE5 and F-UE6 communicate with each other with the help of F-UE7. The F-UE7 is regarded as a mobile relay. In such communication scenario, concerned F-UEs (F-UE5 and F-UE6 in this example) can directly transmit data between each other.
- The network access layer is composed of HPNs and F-APs. The HPNs are deployed to provide system information related to signaling to all F-UEs in the area. However, the F-APs process and forward the data received from the F-UEs. F-APs and HPNs are interfaced with the BBU pool in the cloud computing layer through fronthaul and backhaul links, respectively.
- In the cloud computing layer, the BBU pool is compatible with that of H-CRAN. Moreover, the centralized chasing is also located in this layer.

B. STATE-OF-THE-ART

F-RAN has been studied from various dimensions such as system level performance, architecture, energy consumption, resource allocation, etc. In the rest part of this section, we have tried to the best of our knowledge to review up to date literature related to all these dimensions of F-RAN for 5G mobile communication system. A briefing of this literature review is summarized in Table 7 and its detailed description is provided in the following.

1) PERFORMANCE

The system level performance of F-RAN has been thoroughly explored in [173], [175]–[188]. In [173], the authors study a joint resource allocation and coordinated offloading method for the F-RAN. They have tried to minimize the energy consumption and obtain optimal computational resource allocation for multiple UEs. In [175], the authors explore the trade-off between performance, communication cost, and computing cost in the F-RAN architecture, by using mobile augmented reality as an example. Results obtained by the authors suggest that if the trade-off among these parameters properly handled then the F-RAN can achieve ultra low-latency services. Reference [176] considers the F-RAN with a hierarchical content caching, in which each F-AP is equipped with an individual cache and a part of requests can be responded locally. The authors use stochastic geometry-based theory to derive the average ergodic rate for the content transmission. Moreover, queuing theory has been utilized in order to derive the waiting delay and latency ratio. In [177], the authors develop a system control scheme based on embedded game model for the F-RAN. In proposed scheme, cache placement, spectrum allocation, and service admission algorithms are jointly designed in order to maximize the efficiency of the system. Reference [178] investigates joint mode selection and resource allocation problem in F-RAN supported D2D communication aiming to maximize energy efficiency considering delay and resource reuse constraints.

TABLE 7. A summary of major works related to F-RAN in 5G Mobile Network.

Research Issues	Briefing	References
Performance	Explore system level performance of F-RAN on downlink and uplink, study mechanisms to fulfill QoS requirements of different users, and propose schemes aiming to enhance throughput.	[173], [175]-[188]
Architecture	Investigate physical architecture and topology. Explore key enabling technologies and its implication on F-RAN.	[174], [175], [189]-[193]
Energy	Discover power allocation techniques for the sake of maximization of energy efficiency.	[173], [178], [179], [183], [194]-[196]
Resource Allocation	Investigate efficient allocation of network resources and its assignment to different nodes of the network.	[190], [191], [194], [197]
Surveys	Survey existing literature and contributions related to the F-RAN.	[169]-[172]
CAPEX/OPEX	Develop framework to decrease both CAPEX and OPEX of the F-RAN.	[176]
Mobility	Propose mobility management architecture, mechanisms and feature.	[191]
Spectrum	Address the efficiency of spectrum to improve the performance.	[194]

In [179], the authors study the joint design of multicast beamforming, dynamic clustering and backhaul traffic balancing in the F-RAN. Furthermore, they jointly optimize clustering and beamforming aiming to decrease power consumption, while the delivered service is expected to fulfill the QoS requirements of each backhaul link.

Reducing end-to-end latency is a crucial issue in the F-RAN. Therefore, the authors in [180] are motivated to investigate joint cloud and edge processing design aiming to minimize latency in the downlink of F-RAN. However, the same authors investigate the design of the delivery phase for an arbitrary prefetching strategy in order to populate caches in enhanced RRHS (eRRHs) in [181], [182]. Paper [183] explores information-theoretic analysis of F-RAN. The authors study a latency-centric understanding of the degrees of freedom in high signal to noise ratio regime in the F-RAN limited available resources such as fronthaul capacity, cache storage sizes, power and bandwidth of wireless channel. In [184], the authors propose a dynamic mode selection for the F-RAN using game theory. The authors in [185] propose Markov chain based model in order to analyze the impact of mobile social networks on the performance of edge caching in F-RANs. Furthermore, they analyze edge caching schemes among user equipment in order to minimize bandwidth consumption in the F-RAN. The authors in [186] propose an adaptive resource balancing scheme aiming to maximize serviceability in the F-RAN. The authors of [187] propose a joint distributed computing scheme and a distributed content sharing scheme aiming to achieve ultra low latency by alleviating existing challenges of the fronthaul link. Reference [188] explores a clustering algorithm aiming to maximize throughput in the F-RAN through dynamically determination of the locations of fog nodes. Numerical results presented in the paper show that proposed algorithm increases throughput and decreases latency in F-RAN.

2) SYSTEM ARCHITECTURE

The system architecture of F-RAN has widely been studied in [174], [175], [189]–[193]. Reference [174] studies the network architecture of distributed F-RANs, which enables local data processing, coordinated resource management and distributed storage. Despite exploring system level performance, [175] does also study physical architecture of F-RAN. The authors in [189] discuss hybrid fog-cloud architecture, recent advances in research and system design related issues of F-RAN. In [190], the authors propose NOMA-based F-RAN system architecture for 5G heterogeneous mobile communication system. They further discuss on power and subchannel allocation considering NOMA and edge caching in the proposed architecture. Reference [191] explores system architecture, mobility management, interference mitigation, and resource optimization in F-RAN. The authors of [192] propose a software defined and virtualized RAN with fog computing, where a hierarchical control plane network facilitates fog computing and can be viewed as a centralized F-RAN. Reference [193] addresses system architecture and key techniques for slicing the F-RAN architecture towards 5G mobile communication. Furthermore, the authors present key techniques along with their corresponding solutions, including radio and cache resource management.

3) ENERGY CONSUMPTION

The energy consumption of F-RAN has been studied in [173], [178], [179], [183], [194]–[196]. Reference [194] proposes advanced edge cache and adaptive model selection schemes aiming to improve spectrum efficiency and energy efficiency. The authors subsequently propose radio resource allocation strategies in order to optimize spectrum and energy efficiencies. Paper [195] explores the design of the downlink of multicast in F-RAN and compared soft and hard fronthauling. The authors propose an algorithm that is aimed to

minimize energy efficiency in F-RAN. The authors in [196] survey performance and research challenges for access and fronthaul links in mmWave-based F-RAN. Furthermore, they formulate the optimization problem in the form of non-linear in order to achieve maximum energy efficiency. Moreover, [173], [178], [179], [183] do also explore various dimensions of energy consumption of F-RAN toward 5G mobile network.

4) RESOURCE ALLOCATION

The resource allocation of F-RAN in 5G system is investigated in [190], [191], [194], [197]. The authors in [197] study the trade-off between communication and computing resources in time domain within distributed computing scenario. However, papers [190], [191], [194] shed light on resource allocation mechanisms candidate for the F-RAN architecture of 5G mobile network.

To this end, we have presented the concept, system architecture and related literature of F-RAN proposed for 5G mobile communication system. Despite all advantages that it brings to 5G mobile network, there are still many research challenges and open issues including optimal resource allocation, standards development, theoretical performance analysis, balancing backhaul load, spectrum/energy efficiencies, etc. We provide a detailed description of major research challenges related to the F-RAN in Sec. X.

VIII. A COMPARATIVE STUDY

In Sec. III, we reviewed the evolution of various legacy and traditional RAN architectures. We have discussed that future RANs are reusing existing cellular infrastructure, which reduce both OPEX and CAPEX. Furthermore, we have provided a detailed discussion and comprehensive literature review on C-RAN, H-CRAN, V-CRAN and F-RAN, in Sec. IV, Sec.V, Sec. VI and Sec. VII, respectively. We have pointed out that there are various differences between all these four types of RAN architectures proposed for 5G mobile network.

In this section, we provide comparative analyses of C-RAN, H-CRAN, V-CRAN and F-RAN from various perspectives. In order to provide a well-balanced comparison of above-mentioned 5G RAN architectures, we comparatively analyze their various characteristics provided in Table 8. Our analyses provide insights into how a future-proof RAN architecture can be planned, designed and utilized towards 5G communication system. We discuss key differences among them in the following and compare their major characteristics and architectural comparison in Table 8 and Fig. 18, respectively.

- **Level of Heterogeneity:** The vision of 5G implies a 1000X improvement in the energy efficiency and an average area capacity of 25 Gbps/km^2 (100 times higher in comparison to 4G) [2]. To meet these requirements, various revolutionary approaches and technologies are anticipated including the heterogeneous network. A well-balanced level of heterogeneity has direct influence on capacity, energy consumption and spectrum efficiency of a cellular network. As compared in Table 8 and shown in Fig. 18, the level of heterogeneity in the H-CRAN is very high in comparison to the C-RAN. However, due to higher number of UEs and capacity demands, the V-CRAN and F-RAN do also require high level of heterogeneity to fulfill the requirements of 5G mobile network. It is worth noting that higher level of heterogeneity leads to several interferences, which restricts performance gains and commercial deployments of RAN architecture. Therefore, MNOs are expected to control interferences through advanced signal processing techniques.
- **Deployment of HPNs:** The HPNs are deployed to provide seamless coverage and execute control plane functions. Moreover, the RRHs are used to provide high-speed data rate for the transmission of traffic packets in the user plane. As illustrated in Fig. 18, the HPN in the C-RAN is connected to the core network using backhaul link. However, in H-CRAN, V-CRAN and F-RAN, the HPN are connected to both core network and BBU pool through backhaul links for the purpose of interference coordination.
- **Execution of Network Functions:** Functions such as storage, caching, control, communication, management and CRRM are executed in different locations of proposed 5G RAN architectures. Their qualitative comparison in Table 8 and architectural comparison in Fig. 18 show that above mentioned functions can be executed both in centralized and distributed modes in V-CRAN and F-RAN. Whereas, they can be executed in centralized mode in C-RAN and H-CRAN (with an exception of CRRM in H-CRAN).
- **Decoupling the Control Plane from the User Plane:** Both user and control planes are coupled in C-RAN. Whereas, they are decoupled in H-CRAN, V-CRAN and F-RAN. The separation of control and user planes improves network architecture flexibility, facilitates centralized control logic functions, enhances the performance of network, and ensures easy network slicing for diversified industry applications.
- **Network Complexity:** Adding new elements to the RAN architecture or increased number of user devices introduce additional complexity and overhead, which impact on network performance. We have compared the complexity of four types of proposed 5G RAN architectures in terms of edge networking, fronthaul, BBU pool, RRH and UEs complexity in Table 8. As compared, due to centralization of storage, caching, control, and communication functions, the complexity of edge network is low in the C-RAN compared to H-CRAN, V-CRAN and F-RAN. Fronthaul, BBU pool, and RRHs and UEs complexity can also be found in the table.
- **Data Processing:** In the 5G RAN architecture, the data is processed either in the cloud data center or near to the user device. As for the F-RAN and V-CRAN, the data

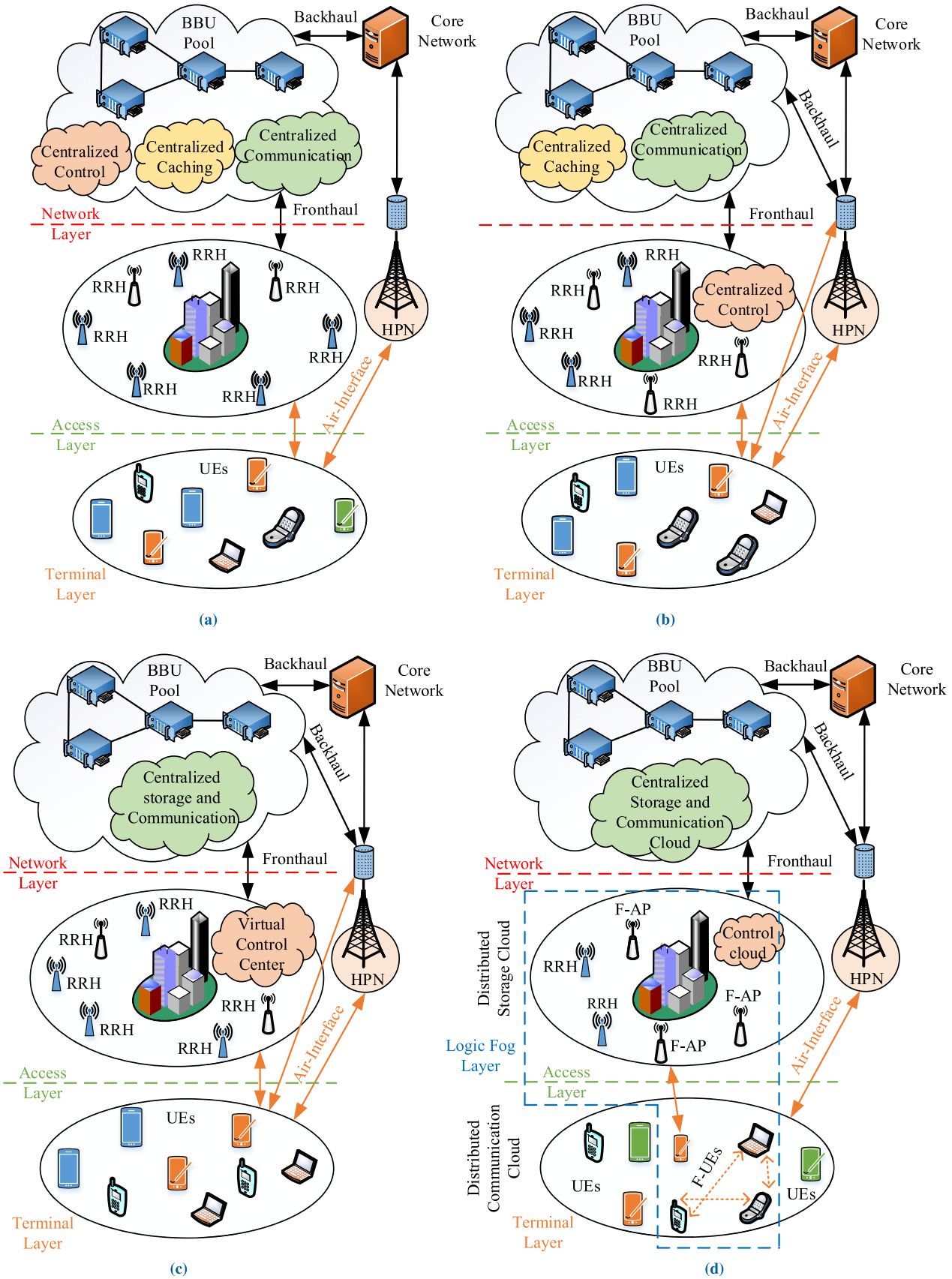

FIGURE 18. Comparison of system architectures of C-RAN, H-CRAN, V-CRAN and F-RAN in 5G mobile networks. (a) System Architecture of C-RAN. (b) System Architecture of H-CRAN. (c) System Architecture of V-CRAN. (d) System Architecture of F-RAN.

TABLE 8. Qualitative comparison of C-RAN, H-CRAN, V-CRAN and F-RAN in 5G mobile networks.

Characteristics	C-RAN	H-CRAN	V-CRAN	F-RAN
Storage	Centralized	Centralized	Centralized and Distributed	Centralized and Distributed
caching	Centralized	Centralized	Centralized and Distributed	Centralized and Distributed
Control	Centralized	Centralized	Centralized and Distributed	Centralized and Distributed
Communication	Centralized	Centralized	Centralized and Distributed	Centralized and Distributed
Management	Centralized	Centralized	Centralized and Distributed	Centralized and Distributed
CRRM	Centralized	Centralized and Distributed	Centralized and Distributed	Centralized and Distributed
Decouple of Control/User Planes	No	Yes	Yes	Yes
Backhaul Interface	CN	CN and BBU Pool	CN and BBU Pool	CN and BBU Pool
Heterogeneity	Medium	Very High	High	High
Edge Networking Complexity	Low	High	High	High
Fronthaul Complexity	Low	High	Medium	Medium
BBU Pool Complexity	High	High	Medium	Medium
RRHs Complexity	Low	Low	Medium	Medium
UEs Complexity	Low	Low	Medium	Medium
Data Processing	Cloud Data Center	Cloud Data Center	Near to Device	Near to Device
Burden on the Fronthaul	High	Medium	Low	Low
Burden on BBU Pool	High	Medium	Low	Low
Burden on CN	High	High	Low	Low
Inter-tier Interference	Low	High	Medium	Medium
Transmission Delay	Long	Long	Low	Low
Latency	High	High	Low	Low
Reliability	Medium	Very High	Very High	High
Energy Consumption	Medium	High	Low	Medium
CAPEX	High	Very High	Medium	Medium
OPEX	High	Very High	Low	Medium

is processed close to the user devices. However, it is processed in cloud data center, when it comes to C-RAN and H-CRAN.

- **Centralization of Network Functions:** Centralization of storage, caching, control, management, etc., brings a significant number of advantages in terms of performance, operation and maintenance to mobile network. However, it does also lead to heavy burden on fronthaul, BBU pool and the core network. Therefore, fully and partially decentralization are introduced in some of the proposed 5G RAN architectures in order to reduce burden on the transport networks as well as BBU pool and core network. The comparative analyses of C-/H-C/V-C/F-RAN in Table 8 show that due to centralization of network functions, burden on fronthaul, BBU pool and CN is high in C-RAN in comparison to H-CRAN. However, it is low in both F-RAN and V-CRAN, thanks to partially and fully decentralization of some of the network functions.
- **Distribution of Network Functions:** The distribution of storage, caching, control, communication, management and CRRM functions have a significant

impact on performance metrics such as transmission delay, latency, inter-tier interference and reliability. As depicted in Table 8, utilizing centralized functions, however, comes at the cost of a potentially large transmission delay, which could worsen the reliability of a system. Whereas, distributedly employment of network functions in the RAN leads to enhance reliability, which is mostly because of the lower transmission delay. The inter-tier interference is directly proportional to the level of heterogeneity. Therefore, the level of heterogeneity is high in H-CRAN in comparison to medium and low levels of heterogeneity in F/V-C RAN and C-RAN, respectively.

- **Total Energy Consumption:** The BS is the main source of energy consumption in a cellular network. According to [6], all base stations of a network operator consume more than 50% of its total energy. Therefore, reducing the energy consumption of BSs contributes to the overall energy consumption of RAN architecture. The power consumption of nodes in the RAN is one of the most important research directions in wireless communication. It determines the intensity of the radiated RF energy into a desired coverage area. Nodes in the RAN are distinguished from each other by, among other parameters, the different transmitter power levels they propagate. The areas where BSs are densely deployed, i.e., urban area should theoretically propagate lower output power than areas with larger distances between BSs, i.e., rural area. As reported in [6], approximately 3% or 600 TeraWatt Hour (TWh) of the worldwide electrical energy is consumed by the ICT sector. Moreover, it is estimated that energy consumption for ICT will grow up to 1,700 TWh by 2030. Therefore, it is important to find new mechanisms of energy consumption in order to reduce energy consumed by the ICT sector and to make telecommunication systems greener. As mentioned earlier, the RAN architecture consume significant amount of energy of a mobile network, thus reducing consumption of cellular access networks will contribute to the energy consumption reductions of entire ICT sector and in particular the mobile network. Due to low network density and centralization of network functions, the energy consumption of C-RAN and F-RAN is medium and the V-CRAN is low. However, due to high level of heterogeneity, the H-CRAN consumes more energy in comparison to other proposed RAN architectures for 5G mobile communication system.
- **Network Expenditures:** The CAPEX and OPEX play vital role during the development and deployment phases of RAN architecture. CAPEX costs are usually high and operators are keen to decrease it by reusing existing infrastructure. The processing capacity of a BS in existing RAN architecture is only assigned to its own UEs and cannot be shared in a large geographical area to be used by the UEs of other BSs. During the day and specially in busy hours the BSs are overloaded in business area while BSs in residential area stay idle, which consume a large amount of power and OPEX. In the H-CRAN, both CAPEX and OPEX are high due to large amount of network nodes, high energy consumption, dense fronthaul/backhaul networks, centralized functions, etc. However, CAPEX and OPEX in C-RAN are high in comparison to V-CRAN and F-RAN and less in comparison to H-CRAN. The virtualization of RAN architecture plays a significant role in decreasing of both CAPEX and OPEX, therefore CAPEX is medium and low in comparison to other proposed RAN architectures. Last but not least, both CAPEX and OPEX are medium in the F-RAN mainly due to edge computing and direct communication among F-UEs in comparison to V-CRAN, C-RAN and H-CRAN.
- **Network Planning and Designing:** The system architecture of 5G RAN defines the location and configuration of network elements and the way these elements are interconnected. It varies from one location to another, from one scenario to another, from one network to another, and from one RAN to another. The density of network elements (such as BSs) in the 5G RAN should be sufficiently enough in order to achieve the targeted RF coverage performance and fulfill the QoS requirements of end-users. If the density of network elements and resources are not sufficiently enough, then, there may be some cases where the mobile operator is not able to provide sufficient network coverage and allocate efficient network resources to the end-users. For example, there may be some locations where the UEs do not have sufficient transmit power to be received by a BS, i.e., coverage is uplink limited. Alternatively, there may be locations where a BS does not have sufficient transmit power to be received by a UE, i.e., coverage is downlink limited. In this example, if a mobile operator designs increasing number of cells for the rural area, it would be an inefficient and costly affair due to low population density. On the other hand, if the mobile operator designs small number of cells for urban area, it would undoubtedly decrease signal propagation which further affects QoS and end-user satisfaction. Therefore, appropriate planning and designing of system architecture and network elements for the RAN architecture of 5G communication system has significant impact on the improving of system level performance, decreasing of CAPEX/OPEX, and reducing the energy consumption.

IX. RADIO ACCESS TECHNOLOGIES FOR 5G

The RAN architecture of 5G mobile network consists of more than a single technology, access method and system architecture. It is a set of selected advanced-technologies, which provides adequate network coverage and enhanced quality of service to various types of end-users and vertical industries. In this section, we discuss and survey key enabling

technologies that are proposed for 5G communication system – more specifically for the RAN architecture of 5G mobile network. These key enabling technologies are expected to enhance spectrum efficiency, decrease CAPEX/OPEX, increase performance, and fulfill diversified requirements of end-users and vertical industries.

Despite various advantages that these key enabling technologies bring to 5G communication system, there is still a number of research challenges that requires further advancement and improvement. We provide a detailed discussion on major research challenges related to all these key enabling technologies in Sec. X.

A. MILLIMETER WAVE

Due to the ever-increasing demand for higher data rate, wireless communications operating in a mmWave frequency band is considered one of the promising solutions to alleviate the current resource bottleneck for future communication systems. With the deployment of mmWave communication networks, the current spectrum bottleneck in conventional microwave LTE systems could be solved with a higher provided bandwidth. Moreover, thanks to the small wavelength, it is possible to equip multiple antenna arrays into limited spaces at mmWave transceivers. As a result, the directional beamforming techniques can be applied to enhance the intended signal strength and mitigate the interference.

The application of mmWave communications, however, is accompanied by several technical challenges in the 5G cellular networks, which requires a novel system to be designed. More specifically, mmWave signals are susceptible to blockages such as buildings or human bodies. Measurement results indicate that there are large scale attenuations in mmWave communication links, which are limited to Line of Sight (LOS) scenarios. In order to predict and evaluate the performance of mmWave cellular networks, new analytical models are needed to be developed.

As illustrated in Fig. 19, the potential deployments of mmWave communications in 5G mobile network are mainly small cell access, cellular access and wireless backhaul. In the 5G RAN architecture, mmWave cells are combined with existing macro cells in order to ensure an always-available connection and enhanced QoS for most of the end-users located in a specific geographic area. The mmWave communication cell offers a coverage range in the order of 100-200 m. Therefore, it is considered as a small cell, which generally occupies the radio spectrum from 30 GHz to 300 GHz, with the wavelength between one to ten millimeters [198].

One of the main characteristics of mmWave communications in comparison to other communication systems (which use lower carrier frequencies) is high propagation loss. The authors of [199], [200] discuss that rain attenuation and atmospheric/molecular absorption characteristics of mmWave prorogation decrease the range of mmWave communications. Therefore, the mmWave communications are usually deployed for short-range/indoor communications

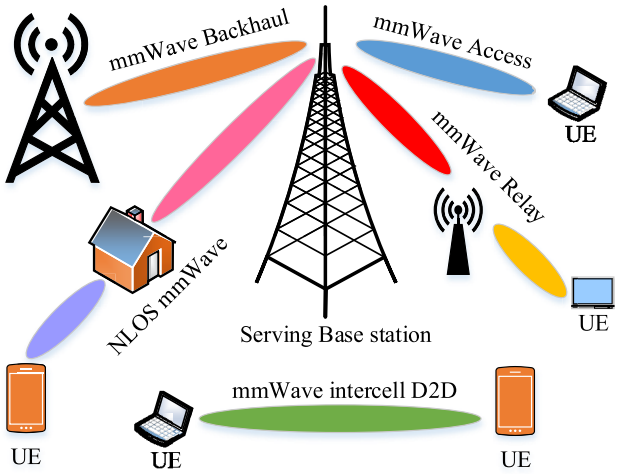

FIGURE 19. Potential deployments of mmWave communication in 5G systems.

such as small cells/backhaul on the order of 200m [201], because, the atmospheric absorption and rain attenuation do not create any additional significant path loss for short-range indoor communication. Directivity is another important characteristic that makes mmWave communication unique. In order for the transmitter and receiver to direct their beams towards each other, several beam training algorithms are proposed, which are available in [202]–[204]. These proposed algorithms are used to decrease beam training time.

Recently, different aspects of mmWave communication have been discussed in several papers including [201], [205]–[213]. In [205], [206], the authors discuss the utilization of mmWave communications in 5G mobile networks. They explore the advantages, feasibility and challenges of the deployment of mmWave communications in emerging cellular networks. Reference [207] investigates the technical challenges of the mmWave communications including atmospheric absorption, phase noise, large-scale attenuation and limited gain amplifiers. The authors do also explore some critical metrics, which can characterize multimedia QoS and furthermore propose a QoS-aware multimedia scheduling scheme that realizes the trade-off between complexity and performance. The authors in [208] discover different deployment strategies of mmWave communication cells in the urban areas. Reference [209] surveys most of the mmWave radio propagation models. The authors deeply discuss different propagation models in terms of line-of-sight probability, building penetration and path loss model. In [210], the authors review mobile networks based on mmWave communications. They carry out a detailed discussion on channel measurements and models, access and backhaul schemes, and MIMO. Moreover, the paper furthermore introduces the standardization and deployment efforts of mmWave communication.

A more in-depth survey on mmWave communication is carried out in [211]. The authors conduct a comprehensive analysis of mmWave communication including four layers - physical layer, MAC layer, network layer and

cross-layer. They provide several use-cases to demonstrate that mmWave communication can satisfy the requirements of future networks. In [212], the authors provide a comprehensive overview of mathematical models and analytical methodology for mmWave cellular systems by taking into account the different types of blockage models and massive antenna arrays, the downlink SINR, and rate distribution were generalized for system-level performance evaluation. This study does also provide a detailed guideline and industrial insights into the future network designation. Reference [213] exploits tools of stochastic geometry in order to obtain a mathematically tractable framework that identifies operating conditions in which mmWave network operates in the inference-limited regime. Last but not least, in [201], the authors present motivation for the deployment of mmWave communication, hardware for mmWave communication measurements, and various types of measurement results in 28 and 38 GHz frequencies in terms of directional antennas at both BSs and end-user devices.

B. MASSIVE MIMO

MIMO is considered one of the key technologies of wireless communication systems for almost a decade. It provides significant increase in capacity and achieves high multiplexing gain. The MIMO systems are divided into two major categories: Point to Point MIMO and Multiuser MIMO. In point to point MIMO, both end-user and BS are equipped with multiple antennas, however, only a single user is served in a single time. In contrast, in multiuser MIMO, a BS is equipped with an antenna array and there are many end-users expected to be served. In order to further increase data throughput and scale up multiplexing gain, the concept of massive MIMO was introduced.

Massive MIMO was proposed for the first time by Thomas Marzetta [214], and is considered as a breakthrough in the 5G mobile communication. It is also known as large-scale antenna systems, very large Multi-User MIMO (MU-MIMO), hyper-MIMO, and full-dimension MIMO systems. Massive MIMO operates in Time-Division Duplex (TDD), however, the downlink and uplink data transmission take place in the same frequency range but in different time domains. As illustrated in Fig. 20, the concept of massive MIMO is to equip a BS with a large number of antennas, which are utilized for transmitting of gigabit-level wireless traffic in order to serve many active users in the same time-frequency. With the deployment of a large number of antennas in a single BS, the challenges of significant gain, high complexity and high cost signal processing technique are emerged. However, the authors in [214] prove that massive MIMO technique is performed optimally. One of the key advantages of the deployment of massive MIMO is the ability to focus the transmitted signal into short-range areas, which leads to improve the performance of system capacity. Massive MIMO does also improve energy efficiency, since tens of antennas located in a BS helps focus energy with an extremely

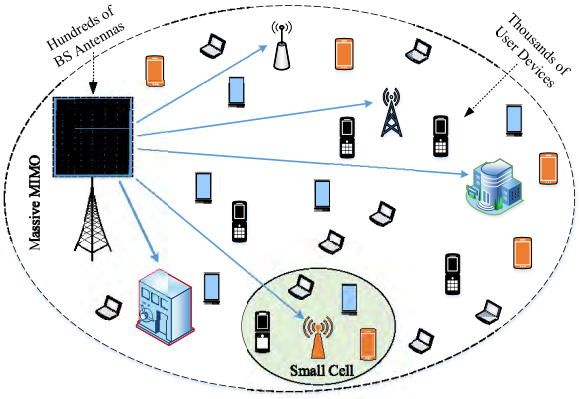

FIGURE 20. The potential deployments of massive MIMO in 5G RAN.

narrow beam on small regions where hundreds of end-users are located.

One of the key advantages of massive MIMO is its integration with other 5G key RATs such as mmWave, Non-Orthogonal Multiple Access (NOMA), heterogeneous networks and D2D communication. Recently, researchers have explored the combination of massive MIMO and NOMA in [215]–[217]. These papers found that NOMA together with massive MIMO improve the probability of network coverage and increase sum rate along with fair radio resource allocation for each end-user. On the other hand, the integration of massive MIMO in heterogeneous network increase throughput by using beamforming technology [218]. The authors in [219] believe that the integration of MIMO and mmWave technologies improve data rates. The small wavelength and high frequency characteristics of mmWave help to design smaller antenna and to pack them together for a realistic massive MIMO technique. Moreover, mmWave can also be used with massive MIMO for point-to-point backhaul links to achieve high throughput. Last but not least, the integration of massive MIMO with D2D improves spectral efficiency. The authors in [220], [221] propose different interference management schemes and resource allocation mechanisms in order to achieve higher performance gains. However, as for the combination of massive MIMO and D2D communication, the trade-off between system complexity and performance has to be taken into consideration.

Massive MIMO has been studied from various perspectives. The authors in [222] explore the energy efficiency of massive MIMO and small cells. They investigate that massive MIMO is more energy efficient if the number of small cells is low, whereas it offers better performance if the number of small cells is high. The authors in [223] explore the energy efficiency optimization problem of downlink of a single cell of massive MIMO systems, which consider both transmit and circuit powers. Reference [224] optimizes data power and user-specific pilot for a given QoS, while the authors in [225] optimize the maximum/minimum spectral efficiency and sum spectral efficiency. The authors in [226] propose a joint power allocation and user association problem for

massive MIMO downlink systems. They investigate to minimize total transmit power consumption, when an end-user is served by a subset of BSs. Last but not least, we recommend two portals (http://massivemimo.eu and http://www.massive-mimo.net/) to interested readers, which provide a list of related technical/overview papers, highlight most recent literature, and share experimental/scientific results related to massive MIMO.

C. DEVICE-TO-DEVICE COMMUNICATION

In D2D communication, the data traffic of closely located devices is directly routed and is not necessarily needed to traverse through RAN or CN. Recent studies have shown that due to short distance among pair mobile devices and direct communication, the D2D communication is expected to enhance performance, increase energy efficiency, and decrease delay in mobile network. An example of D2D communication along with cellular communication is depicted in a heterogeneous cellular layout in Fig. 21, where pair devices directly communicate with each other, and the cellular communication devices communicate with the eNB.

The D2D communication is categorized into in-band and out-band. In the in-band D2D communication mode, the direct communication among devices takes place in a licensed spectrum allocated to the cellular operators. The in-band D2D users access the licensed spectrum in two modes: the dedicated mode (overlay or orthogonal mode) and the shared mode (underlay or a non-orthogonal mode). In the out-band D2D communication mode, the direct communication among devices takes place in unlicensed spectrum adopted by other wireless technologies such as Wi-Fi or Bluetooth.

D2D communication has exclusively been studied from various dimensions. So far, more than two hundred technical, overview and survey papers have been published related to D2D communication. Reviewing of all these papers in this subsection is not possible due to limited space and considering main objectives of this survey paper. However, we highlight only those survey papers, which review D2D communication in 5G mobile networks from different perspectives. In this way, on the one hand we will thoroughly address D2D communication in 5G systems, and on the other hand, we will also give corresponding references to interested readers to explore more about a specific dimension of D2D communication in future networks.

Reference [227] provides a review of existing literature on D2D communication from spectrum perspective. The authors categorize available literature into two major classes of D2D communication, the in-band D2D communication and the out-band D2D communication. The paper further explores open research challenges and future directions. In [228], the authors review D2D communication from protocols, resource allocation techniques, mobility management, algorithms, and architecture perspectives. Existing open issues and future research directions are also provided. Reference [229] explores available literature related to interference management, radio resource management and optimization, performance evaluation, applications and services, implementation, standardization activities, open issue, and future research directions related to D2D communication.

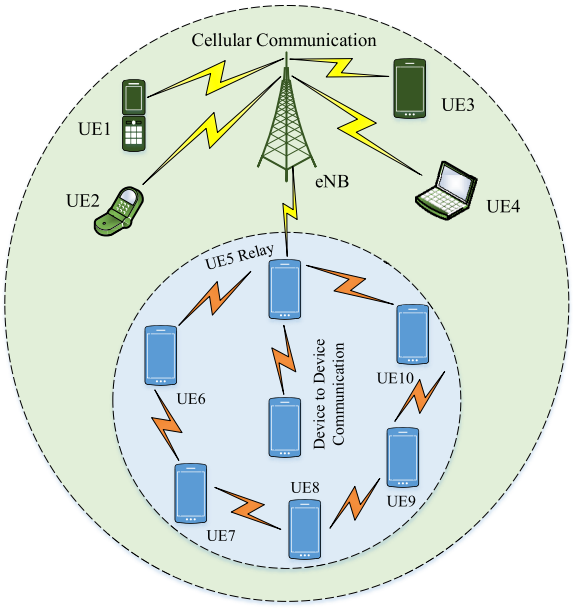

FIGURE 21. The deployment of D2D communications in cellular networks.

With the deployment of D2D communication in the cellular network, the interference management between D2D users and cellular users is one of the key challenges that needs to be taken into account. In [230], the authors survey existing literature on interference management of D2D communication in cellular network. The paper furthermore classifies and compares all interference management techniques. The study has shown that existing D2D techniques need to be further optimized in order to fulfill the requirements of 5G mobile communication system.

Routing in multi-hop cellular D2D networks is a critical issue that needs to be carefully designed. Because, network fragmentation, dynamic network topology, and node mobility are the main characteristics of multi-hop networks, which do not exist in the traditional cellular network. Reference [231] surveys this specific dimension of D2D communication. The authors provide a comparative analysis of routing schemes and discuss open research directions of routing in multi-hop cellular D2D communication networks. In [232], the authors explore applications and socially-unaware technical challenges of D2D communication. The paper surveys and categorizes literature related to socially-aware D2D communication, and subsequently, discuss existing open research challenges and future directions. Reference [233] surveys available literature of D2D communication in the context of 3GPP LTE/LTE-A. To address several dimensions of D2D communications, the authors classify related papers into D2D management, D2D scenarios, and D2D RRM categories. In [234], the authors provide a review of existing state-of-the-art solutions for channel measurements and modeling.

Security and privacy are important aspects in all types of wireless communication, and especially when it comes to the deployment of D2D communication in cellular networks. Both of these dimensions of D2D communication have widely been surveyed in [235]–[238]. Reference [235] classifies literature related to security of D2D communication in physical, MAC, network, and application layers. The authors believe, layer-based security approach offers a comprehensive understanding, which helps to improve security and design the protocol of D2D communication. The authors in [236] surveys state-of-the-art solutions related to architecture and security of D2D communication. Despite survey, the paper proposes an architecture, which is claimed to overcome deployment challenges of D2D communication. The authors furthermore propose a solution concept, which is aimed to enhance security of D2D communication using IP security. The paper does also survey different types of security attacks that D2D faces with. In [237], the authors investigate security architecture of D2D communication in the context of 3GPP LTE framework. The paper does also explore and survey potential security threats, security requirements, and existing security solutions in D2D communication. Reference [238] reviews state-of-the-art solutions for enhancing privacy and security in D2D communication. The authors categorize all proposed security solutions of D2D based on cryptographic design, pairing and discovery, and distributed algorithms. The paper further classifies state-of-the-art related to privacy of D2D communication into two categories: device privacy and network privacy. The device privacy is covered from access control, privacy policy, application analysis, data leakage, and mobile operating systems perspective. However, the network privacy is surveyed from anonymity, trust, access control, communication, storage access, private proximity testing, and location privacy.

D. MASSIVE MACHINE TYPE COMMUNICATION

The scenario of mMTC is defined to concern emerging concepts such as like sensor networks and IoT, which require an ultra high traffic density and a huge number of connectivities. In traditional telecommunication networks, the network services were limited only to smart-phones, however, with the deployment of 5G, the network services are going to be gradually introduce to various devices around us such as smart watches, smart sunglasses, etc. Therefore, the 5G is also characterized with full interconnection of all things, and the mMTC is widely accepted as an indispensable part of 5G. Through the Internet of Everything (IoE) concept, the 5GPPP expects the 5G network to connect up to trillions of IP-based ''things'' by 2020 [239], [240]. For instance, the 5G network will be a systematic part of future smart cities in which the services and applications of 5G will have a huge impact on energy management, water management, smart networked households, smart/intelligent vehicles, tele medicine/surgery, public safety, education, traffic management, time-critical applications that require an immediate reaction, etc. As a result, this will make the planning, management, and maintenance of next generation of telecommunications networks expensive and complicated. In comparison to classical handheld Human-Type Communications (HTC), the mMTC exhibits unique characteristics of [241]:

- Huge number of devices;
- Small data packet size, potentially down to a few Bytes;
- Uplink-dominated transmissions;
- Low user data rates;
- Sporadic user activity, including periodic and(or) synchronized transmissions;
- Strong constrains on device complexity and power efficiency.

So far, the mMTC has attracted significant attention from both academia and industry. There are various survey and review papers available, which provide detailed discussion on the mMTC from different perspectives such as application requirements along with their associated communication technologies in [242], access management techniques in [243], radio resource management in [244], economic considerations for facilitating the deployment of MTC in [245], and surveys on mMTC in perspectives of service requirements and technical challenges are available in [246], [247].

The demand for dense radio access over shared medium with low latency is a problem with long history. In case of traffic congestions, frequent Random Access (RA) collisions may lead to a significant delay over the RACH, and therefore must be efficiently suppressed. A variety of solutions for RA collision control have become available since the early versions of LTE standard, including Access Class Barring (ACB), Prioritized RA, MTC-Specified Backoff, RACH Resource Separation, Dynamic RACH Allocation, Pull-Based RA, Self-Optimization Overload Control RA, Code-Expanded RA, Spatial-Grouping, Guaranteed RA and Non-Aloha-Based RA. These legacy methods have been deeply reviewed in performance and implementation through comparative studies [248], [249].

To cope with the new requirements of 5G mMTC scenario where the UE density is significantly further increased, novel methods have been reported. For example, it was proposed in [248] to apply a reinforcement learning algorithm to cognitively address UEs to different BSs so that the peak PRACH load can be distributed over BSs. Since 2015, a new concept has been arising to a recent popularity, which deploys D2D technology to enable some UEs working as data aggregators for other devices [250]–[253]. In such way, the RA request density in mMTC scenario can be efficiently reduced, and therefore so does the collision rate. Additionally, making use of the UE context information including channel gain and battery level, an optimized assignment of data aggregator can lead to a significant improvement in UE energy efficiency [254]. Another D2D-based approach, as an alternative for MTC among local devices in the same area, relies on D2D interfaces between UEs (such as PC5ah) to set up local ad-hoc networks without any relay over BS, so that RAN congestions can be avoided [255]–[257].

E. MULTI-ACCESS EDGE COMPUTING

In order to cope with explosive growth of data traffic that is associated with a wide plethora of emerging applications and services, a new architecture, which is called the Multi-access Edge Computing (MEC) was proposed. MEC is a new paradigm, which provides cloud computing services at the edge of the radio access network, offering an ultra-low latency environment with high bandwidth and real-time access to radio and analytic. MEC enables computationally intensive and delay-sensitive applications to be executed in close proximity to end-users, thereby enriching user's satisfaction, improving QoE, and utilizing more efficiently the mobile backhaul and core networks. From business perspective, MEC introduces flexible and multi-tenant environment, which allows authorized third-parties to make use of the storage and processing capabilities.

Key technologies that enable edge computing are SDN, NFV, Information Centric Networking (ICN) and network slicing. The ETSI standard on MEC [258] plays an active role by launching the MEC ISG. The objectives of ISG are to accelerate the adoption of edge technology by developing set of specifications in the form of deliverables, including service scenarios, requirements, architecture and API specifications.

There are various promising scenarios and use-cases in 5G communication systems that benefit from the concept of edge computing such as Computation offloading, IoT, content delivery and caching, AR and VR, video streaming and analysis, connected vehicles, and mobile big data.

Computation offloading is considered a significant use-case of MEC, which improves energy efficiency and speeds up the process of computation – especially the delay-sensitive applications. Computation offloading techniques have attracted extensive research efforts, which are reported in [259]–[264]. For example, [259], [260], [262], [264] exploit the computation offloading mechanism for enhancing the capability of mobile devices, while improving energy efficiency. However, papers such as [261], [263] study execution delay introduced by computation offloading. Besides, an exclusive overview on computation offloading is provided in [265]. The authors describe three important design challenges in computational offloading to MEC: offloading decision, mobility management and computation resource allocation.

The MEC can be utilized for IoT applications and services by facilitating storage and computational related resources close to the data sources [266]. For instance, the MEC can be exploited to ensure that end-user's requests are processed faster or to reduce IoT data and signaling. The architecture for edge IoT proposed in [267] suggests a hierarchy of edge cloud platforms in order to process IoT data. The proposed architecture does also collect and analyze raw IoT data, which are subsequently delivered to the remote cloud server. The results of a research in [268] investigates the freshness of IoT data suggests that the load of the network can be significantly reduced for highly requested data. The authors in [269] explore VM placement and task distribution in edge cloud considering end-user QoE and CAPEX. Moreover, [270] elaborates a cloud-based controller known as IoT-Cloud that enables the developers to create scalable sensor-centric applications.

The responsibility of the edge network can be further increased by allowing distributed caching with hosting Content Delivery Network (CDN) at the edge of the network. MEC-CDN are capable of reducing stress from core network [271]. The authors of [272], [273] do also confirm such outcomes through the results of their research presented in their work. Moreover, the research work presented in [271], [274], [275] suggest various proposal related to the edge-assisted distributed video caching and streaming to increase the QoE of end-users and QoS of the 5G mobile network.

One of the promising applications of MEC is to support delay-sensitive computing such as AR and VR technologies [276]. The results found in [276]–[278] show that the offloading of AR application tasks to the edge reduces latency and improves energy efficiency. Reference [279] does also present the performance evaluation of computing task offloading to edge server for AR application. The results show that energy consumption is reduced and the latency is decreased due to the offloading of tasks to the edge.

In Sec. I and Sec. II, we discussed that nearly three fourths of mobile data traffic of 5G communication system is video traffic. The deployment of edge caching facilitates essential operations of video stream analysis including the detection of object, and the classification of wide-range of applications such as face recognition, home security surveillance, and vehicular license plate recognition [280], [281]. Considering the fact that video analysis algorithms need high computation resources, thus it is preferable to offload video analysis tasks to the MEC in order to enhance energy efficiency (such as increasing battery life) of the user devices, reduce latency and provide higher data-rates. The performing of video analysis close to the end-user devices (edge network) does not only decrease energy consumption and reduce the latency, but also avoids network congestion in the 5G mobile network caused by video stream uploading [276], [280], [281].

MEC plays a significant role in the digitization of automotive industry such as self-driving and autonomous cars. It can also be exploited for Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communication in order to provide services for the applications that require ultra low-latency. The deployment of MEC alongside the road sends and receives crucial information that alleviates V2V and V2I communication such as traffic jam, warning from other vehicles, and the presences of pedestrians and bikers [282]. Therefore, the deployment of MEC in automotive industry improves road safety due to the combination of cameras mounted on the vehicles with MEC video analytics at the edge of the network [282]. Recently, a study on the architecture of vehicles connectivity through edge-cloud platform reports a number of V2V and V2I on-board services [282]. Moreover, the authors in [283] discuss that vehicles, which are connected via the distributed edges collect, process, and

analyze real-time data from sensor devices that are installed ubiquitously. Subsequently, the data is transferred to the cloud data center for further processes.

The deployment of MEC close to the end-user facilitates big data analytics and offers low latency services with higher bandwidth [284] [285]. For example, in [284] a framework is introduced to further exploit the collaborative processing between edge and cloud computing for big data analytics. In this context, the authors suggest that IoT data can be collected in the edge of the network and after the analysis, the result can be sent to the core network.

F. NETWORK FUNCTION VIRTUALIZATION

Most traditional mobile networks are filled with large amount of network functions, which are coupled to dedicated hardwares. This mechanism brings many challenges to the network such as difficulty in service provision and network management. Moreover, deployment of dedicated hardware and designing of protocols based on dedicated hardwares are both expensive and time consuming. The NFV is considered as a key enabler that decouples network functions from dedicated hardwares and realizes them in the form of software, which is called Virtual Network Functions (VNFs). The NFV introduces many advantages to the telecommunication networks such as simplify network management, decrease CAPEX/OPEX, reduce energy consumption, enhance flexibility of service provision, and so on.

In October 2012, some of the world largest telecommunication operators and vendors such as American Telephone and Telegraph (AT&T), British Telecom (BT), Deutsche Telekom (DT), NTT, DOCOMO, Telecom Italia, Telefonica, and so on agreed to establish the first Industry Specification Group (ISG) within ETSI in order to define NFV in telecommunication networks [286]. Since establishment of ISG on NFV within ETSI, the number of members increased to more than 300 including 40+ world largest telecommunication service providers. All these members are closely working together on deployment of NFV in telecommunication networks and development of standards for NFV.

The NFV architectural framework is illustrated in Fig. 22, which is proposed by the ETSI [286]. The proposed architecture consists of three parts: the Network Functions Virtualization Infrastructure (NFVI), Virtualized Network Function (VNF), and NFV Management and Orchestration (NFV M&O). The NFVI corresponds to the data plane, which is used to provide virtual resources in order for the VNFs to be executed. The VNF is the software implementation of a network function and has the capability to run over the NFVI. The VNF corresponds to the application plane and consists of various types of VNFs, which are considered as applications. The NFV M&O part of proposed architecture corresponds to the control plane. It is responsible for the orchestration and lifecycle management of hardware and software network resources, which are used to support lifecycle management and infrastructure virtualization. Moreover, the M&O has responsibility to build connection among different VNFs

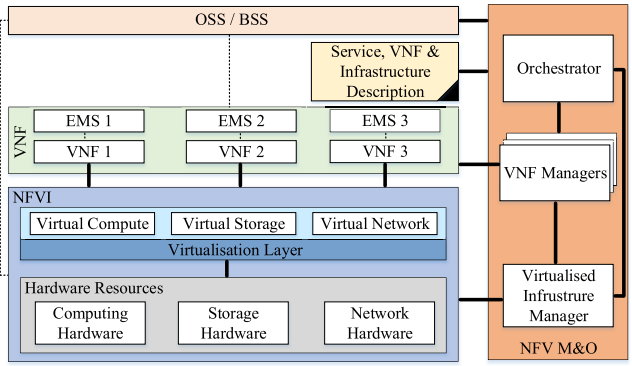

FIGURE 22. The NFV architectural framework.

and interacts with the OSS/BSS landscape. This feature of M&O allows NFV in order to integrate with existing network management landscape. The NFV system is driven by a set of metadata that consists the requirements for service, VNFs, and infrastructure in order for the NFV M&O to act accordingly. All these descriptions, VNFs, infrastructure, and services are provided by various industries.

The 5G communication network is filled with large amount of functions. All of these functions are faced with functional and architectural challenges. In order to effectively process and adopt these function in the 5G network architecture, the NFV is one of the key enablers to achieve this goal. The NFV provides high flexibility to adopt to various scenarios, requirements, and use-cases of 5G and beyond communication technology. In the 5G RAN architecture, there are several functions related to user plane and control plane that need for virtualization. The virtualization of functions in 5G and more specifically in 5G RAN decreases energy consumption and lowers footprint through its dynamic infrastructure resource allocation and traffic balancing features.

There are many research papers available that investigate the integration of NFV in different traditional telecommunication networks. For example, the authors in [287], [288] investigate possible deployment of NFV in optical communication networks. Papers such as [289]–[291] study the deployment of NFV in different aspects of 5G mobile network. Moreover, there are also some papers that study the integration of NFV with SDN such as [292]–[294], which attracts significant attention due to its high complementary features. However, the integration of NFV with traditional systems is not easy and requires more studies in order to solve compatibility issues.

There are many survey and review papers available, which explore various aspects of NFV such as [9], [295]–[300]. In [295], the authors provide an overview of NFV and also discuss most popular algorithms related to VNF in terms of scheduling, migration, VNF placement, chaining and multicast. Moreover, existing open challenges and future research directions are also highlighted. Reference [296] surveys state-of-the-art that leverages NFV and SDN in the evolved packet core network architecture. The authors categorized related

literature in four categories, architectural approach, technology adoption, functional implementation, and deployment strategy, and discussed all these four dimensions of NFV/SDN deployment in core network architecture. In [9], the authors provide a comprehensive review of NFV and furthermore explore the relationship of NFV with the SDN and cloud computing. The paper also deals with standardization activities, NFV related projects, deployment, potential use-cases, and commercial products. Reference [297] surveys radio resource allocation of NFV and furthermore presents main research challenges related to resource allocation in the NFV. Reference [298] investigates the deployment of NFV in software-defined NFV architecture. The authors present software-defined NFV architecture and describe the relationship between SDN and NFV. In [299], the authors survey recent papers on the deployment of NFV along with the advantages and disadvantages of NFV and SDN. Last but not least, [300] explores NFV and furthermore covers system requirements and its framework. The author discuss potential NFV use-cases and highlight open challenges and future research directions.

G. SOFTWARE DEFINED NETWORKING

Legacy telecommunication networks are integrated vertically, where control plane and data plane are paired together. The existing integration approach of traditional IP networks is complicated, difficult to manage and hard to configure. The SDN is considered as a key enabler of 5G and beyond communication technologies, which breaks traditional integration approach of telecommunication networks and decouples control plane from the data plane. The SDN has the ability to centrally partition network, change the traffic flow, and provide application level quality of service. The decoupling of data and control planes brings a number of advantages to a communication network such as simplification of network management, service management, control management, and it become easy to program the applications. However, due to some challenges such as budget constraints, fear of downtime and so on, many telecommunication operators are not willing to fully deploy SDN, instead, they are willing to chose partially deployment of SDN for the design of their networks. The main idea behind partially deployment of SDN is to place limited number of SDN related hardwares among existing telecommunication network devices, which is also called hybrid SDN network.

The SDN architecture is depicted in Fig. 23, which consists of three layers: infrastructure layer, control layer and application layer. The infrastructure layer is the bottom layer, which is composed of all network devices and hardwares related to SDN supported system. In contrast with traditional network, devices in the SDN network do not have control over functions, and they act just as forwarding devices. As shown in the architecture, the infrastructure layer interacts with control layer via control data plane interface. The control layer is composed of multiple SDN controllers. All intelligence related to the network is logically centralized in this layer.

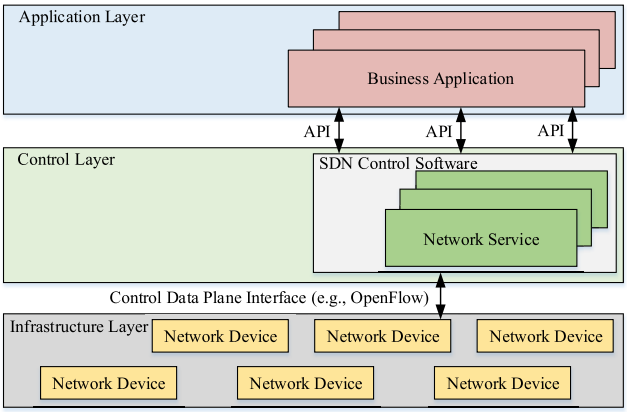

FIGURE 23. Software Defined Networking architecture.

The SDN controllers are responsible to manage all virtual and physical network resources. As illustrated in Fig. 23, the SDN controllers have direct control over all elements of data plane through Application Programming Interface (API). The application layer is the upper layer, where service providers, network operators and application developers directly operate the network to meet their business demands such as bandwidth, traffic, access control, QoS, energy usage, etc.

The 5G mobile network infrastructure is based on SDN, which is supposed to be managed dynamically and on the real-time needs. This will help operators to provide efficient communication between applications and services in the cloud, and end-users. Moreover, the deployment of SDN in 5G mobile network increases efficiency in allocation of radio resources through centralization and seamless mobility over diverse radio access technologies. There are many research papers and proposed solutions available, which discover SDN in the context of 5G network and more specifically in the 5G RAN such as [14], [301], [302]. In paper [14], the authors model an architecture, in which SDN controller adjusts bandwidth for each Radio Access Points (RAPs) and BBU dynamically. In the proposed architecture, SDN controller has the responsibility of management and selection of routes for all RAN and CN connections. In [301], the authors propose a novel architecture in order to leverage SDN in the 5G wireless network, which is called SoftAir. The paper discusses solutions and existing challenges regarding proposed architecture. The authors in [302], propose a multi-tiered cloud controller scheme and event processing mechanism for Software Defined Wireless Network (SDWN) architecture of 5G mobile communication.

Despite above discussed technical papers, there are several survey papers available, which study SDN from different aspects such as [303]–[310]. In [303], the authors provide a comprehensive survey on SDN infrastructure, network programming languages, network applications, southbound and northbound APIs, and standardization activities. Reference [304] surveys recent research papers on SDN,

its features, and its potential benefits. The authors thoroughly describe all three layers of SDN network architecture. The authors in [305], [306] investigate all those papers, which are focusing on three layers of SDN network architecture. However, [305] surveys literature related to the interfaces between aforementioned three layers. Reference [307] explores literature on SDN advancement over traditional network, SDN technologies, functional architecture, OpenFlow standards/protocols, and various activities that are going on to standardize different aspects of SDN. The authors in [308] explores design principle, architecture, application, and potential driving forces of SDN deployment in the future. In [309], the authors survey related literature on different aspects of SDN. Moreover, the paper focuses on discovery of network model from both architecture and protocol perspectives. The authors discuss various available solutions, which address scalability, elasticity, dependability, reliability, high availability, resiliency, security, and performance concerns related to the SDN. Paper [310] discusses literature related to programmable networks, specifically the SDN. The authors present the architecture of SDN supported system, OpenFlow standard, and alternatives for deployment of SDN-based protocols and services. It is worth noting that if interested readers are looking for open issues, ongoing research efforts, and future directions related to SDN, then we recommend references such as [303], [304], [306], [310], where above mentioned aspects of SDN are thoroughly discussed.

H. NETWORK SLICING

As discussed in Sec. I, the 5G communication system is expected to provide services to various vertical industries such as medical care, manufacturing, automotive, etc. Network slicing is one of the appropriate 5G technologies to meet the requirements of vertical industries. It allows operator to partition network in a structured, elastic, scalable and automated manner. Each of the use-cases and applications demands its own network slice that consists of independent functions, requirements and characteristics [311]. For example, a slice may be dedicated to Critical-Machine Type Communication (C-MTC) such as remote surgery, which is typically characterized by high reliability, ultra-low latency and high throughput. Another network slice may be specified to support water meters reading, which requires a simple radio access procedure, small payload volume and low mobility. Furthermore, the eMBB services may require a separate slice, which is characterized by a large amount of bandwidth in order to support high data rate services such as HD video streaming.

Deployment of network slicing enables the operation of multiple logical networks over a single shared physical infrastructure using SDN/NFV in order to reduce total cost, decrease energy consumption and simplify network functions in comparison to one network for different use-cases/business scenarios. In a slice-based network, each slice is provided with its own specific characteristics and will be considered as a single logical network. In this way, infrastructure utilization and resource allocation will be much more energy and cost efficient in comparison to traditional network.

Implementation of network slicing over 5G communication system arises many technical challenges, which are expected to be solved. There are also some business and economic issues (e.g. total cost, revenue, etc.) that need for significant optimization and re-designing in order to cope with emerging network architecture. On the other hand, the demand for broadband multimedia services has been increasing explosively. With this ongoing trend, the revenue of MNOs will soon be exceeded by CAPEX and OPEX required to operate the infrastructure. Therefore, total cost, expected revenue, and resource allocation in the context of network slicing are seemed to be interesting research topics, which need for further discovery.

Fig. 24 shows the system architecture of network slicing [312]. The architecture consists of CN slices, RAN slices and radio slices. Each slice in CN is built from a set of Network Functions (NFs). Some NFs can be used across multiple slices while some are tailored to a specific slice. There are at least two slice pairing functions, which connect all of these slices together. The first pairing function is between CN slices and RAN slices, and the second pairing function is between RAN slices and radio slices. The paring function routes communication between radio slice and its appropriate CN slice. The pairing function between RAN and CN slices can be static or semi-dynamic configuration in order to achieve required network function and communication.

Network slicing has exclusively been studied in literature. For instance, the authors in [313] explain a detailed end-to-end framework of network slicing implementation in 5G communication system. The paper deals with the deployment of vertical and horizontal slicing over the air-interface, RAN and CN. It further focuses on how to horizontally slice both computation and communication resources to/from virtual computation platforms in order to improve scalability, enhance device capability and increase end-user experience.

Moreover, [314]–[316] focus on the deployment of network slicing in the RAN architecture. The authors in [314] analyze network slicing in multi-cell RAN in order to support radio resource splitting among various slices. The paper further proposes four types of RAN slicing approaches along with their detailed comparison. However, the authors in [315] explain how network slicing impacts various aspects of design and functions of RAN architecture of 5G mobile network. The paper thoroughly covers RAN requirements for network slicing implementation. In [316], the authors provide a comprehensive discussion on deployment of network slicing in H-CRAN aiming to improve throughput through computation and communication resource sharing.

Furthermore, the authors in [317] address network slicing related concepts, i.e., resource allocation, virtualization technologies, orchestration process and isolation function. The paper provides a comprehensive discussion on SDN and NFV along with a deployment use-case (which considers network slicing using both NFV and SDN integration).

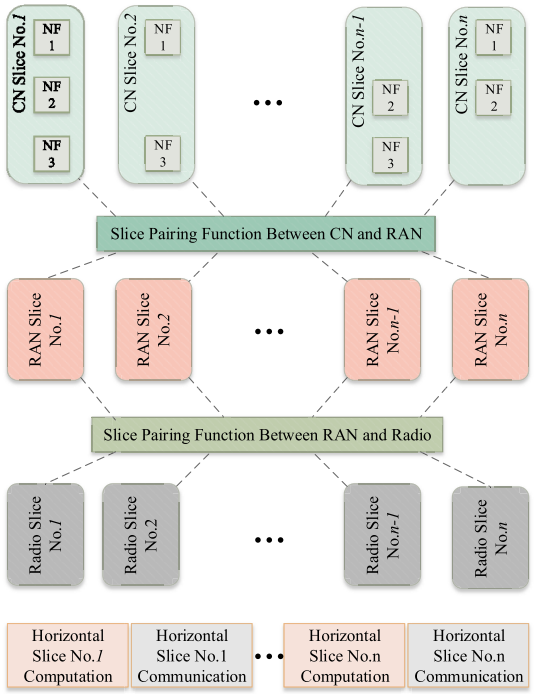

FIGURE 24. Network slicing system architecture.

The authors further demonstrate existing challenges and future research directions of network slicing implementation in 5G communication system. Moreover, a comprehensive survey on architecture and further research directions of network slicing is also available in [318].

The Next Generation Mobile Networks (NGMN) introduced Network Virtualization Substrate (NVS), which allows infrastructure provider to control resource allocation of each of the virtual instances of an enhanced Node B (eNB) before customization of scheduling of each virtual operator within allocated resource. On the other hand, a heuristic-based admission control mechanism is introduced in [319], which dynamically allocates network resource to various slices in order to increase end-user satisfaction considering specific requirements of each of the slice.

One of the main objectives of network slicing is to optimize profit modeling of traditional telecommunication networks. In order to increase overall revenue and decrease total network expenditure, a comprehensive study of business and economic dimensions of network slicing is required. The authors in [320] analyze profit generated by different slices over the same network infrastructure and furthermore model network resource management. Meanwhile, [321] deals with the designing of an algorithm that allocates requests of network slices, which maximizes total revenue of network infrastructure provider.

X. OPEN RESEARCH CHALLENGES AND FUTURE DIRECTIONS

In this section, we explore several open research challenges related to various 5G RAN architectures (C-RAN, H-CRAN, V-CRAN, and F-RAN), key enabling technologies for 5G systems (MEC, SDN, NFV, and network slicing), RATs of 5G mobile network (mmWave, massive MIMO, D2D communication, and mMTC), and transportation networks of 5G RAN architecture (Backhaul and Fronthaul). We further align existing challenges for future research consideration.

A. MAJOR RESEARCH CHALLENGES OF C-RAN

- **Security:** Existing security frameworks cannot defend all possible treats and attacks against the C-RAN. Therefore, a universal security framework is essential to fulfill the requirements of 5G C-RAN. Among others, privacy preservation, trust management, security of spectrum resource management and physical layer security are considered to be the most crucial challenges. Based on these four major research challenges, we believe, there is a number of potential research directions including the design of: an efficient and secure authentication mechanism, a comprehensive security framework, a mechanism to allow various operators in order to share their resources among each other in a trustworthy way, a privacy preservation mechanism, a system that builds trust among end-users and service provider, and a secure virtualized mechanism for virtualized BBU pool.
- **SDN Deployment:** The deployment of SDN in C-RAN brings a number of advantages to 5G mobile network. However, it does also lead to two major research challenges: First, the placement of centralized controller needs optimization, because, the location of the controller play crucial role in latency, QoS, and other network performance parameters. Second, network operators are faced with the scalability challenge in the C-RAN, which is limited by service capability of SDN controller.
- **Cooperation among BBUs:** All BBUs, which are located in the same pool, require cooperation among each other in order to share end-users' related data, scheduling and channel feedback collection. In 5G mobile network, C-RAN architecture will be equipped with a huge number of BBUs that are expected to be located in the same pool requires high cooperation. Despite many advantages that cooperation among BBU brings to the 5G C-RAN, there is also a number of challenges including end-user privacy, high bandwidth and low latency; which need to be tackled.
- **Cells Clustering and BBUs Assigning:** Efficient assignment of BBU pool and optimal clustering of cells in an effective way to minimize overhead and maximize gain are still open research challenges. There could be a scenario, where, one BBU pool may achieve a high number of channels, meanwhile, it may reduce fronthaul delay. However, there may be another BBU pool, which supports many distributed geographical locations in different places in order to consolidate them with a single BBU. Therefore, we need to find an efficient mechanism in order to cluster cells and assign BBU in 5G C-RAN.

B. MAJOR RESEARCH CHALLENGES OF H-CRAN

• **Energy Consumption:** Ultra dense deployment of RRHs and micro BSs in H-CRAN, on one hand improves capacity, on the other hand consumes too much energy. Existing literature related to the optimization of energy efficiency assumes that transmission power of every node/network is fixed or slightly increases/decreases. Therefore, energy harvesting in the H-CRAN is one of the promising research challenges that needs further study into many directions including the adaptation of transmission power of node/network to the packet traffic, radio channel fading, the QoS of users, etc.

• **Backhaul Links:** Both intra-CoMPs and intercell CoMPs need a large amount of signaling over backhaul links in order to mitigate interference among macro and micro BSs. This leads to increase the capacity of the backhaul links. Therefore, the H-CRAN requires low latency and extremely high capacity backhaul network to overcome existing capacity challenge.

• **Allocation of Workload between Smallcells and Macrocells:** The RRHs and small cells are usually deployed indoor and in hot traffic zones in order to provide enhanced quality and reliable communication to the end-users of H-CRAN. This will increase the amount of data to be transmitted between RRHs/smallcells and the macrocells. Therefore, an efficient mechanism to allocate workload between RRHs/smallcells and macrocells is needed to be researched in order to take a large amount of traffic demands into consideration.

C. MAJOR RESEARCH CHALLENGES OF V-CRAN

• **Overhead and Latency:** Network functions are executed on top of a hypervisor in NFV system architecture, this characteristic of network virtualization introduces additional overhead in V-CRAN, which leads to network performance degradation. Therefore, optimization of legacy hypervisors and deployment of new virtualization technologies are needed in order to minimize overhead and latency.

• **Network Isolation:** With the introduction of NFV and SDN in V-CRAN, network is isolated. Thus, both physical and virtual resources are shared among different operators. Different types and levels of configuration, customization, and deployment in the topology of virtual network should not affect and interfere in any part and service of coexisted operators. Network isolation is a challenging task in wireless networks due to the broadcast nature of wireless communication and propagation of radio waves. For instance, a small change in a cell configuration may introduce a high level of interference to its neighboring cells. Therefore, special attention should be paid during configuration, customization and deployment of both physical and virtual resources.

• **Resource Allocation:** Efficient resource allocation and utilization is another crucial challenge of V-CRAN. Some of the characteristics of wireless communication network such as the availability of spectrum, roaming, device mobility, and the differentiation of uplink and downlink channels make the resource allocation more complicated in comparison to wired networks. Therefore, researchers are expected to propose efficient resource utilization mechanisms for V-CRAN in order to maximize the profit of mobile network and increase the QoS of end-users.

• **Network Slice Management:** Network slice management is a major research challenge in V-CRAN. Every single network slice should be created and scaled dynamically based on its QoS requirements. Those algorithms that are utilized for the allocation of network resources should adopt different strategies depending on the QoS requirements, size, and application of network slice. Therefore, dynamic characteristics of 5G network slice need to be taken into account while designing and deployment of network resource allocation algorithms in V-CRAN.

D. MAJOR RESEARCH CHALLENGES OF F-RAN

• **Integration of SDN:** The F-RAN decouples control and user planes. The data forwarding flow in SDN is mainly at IP layer. It is still not straightforward that how to combine MAC and physical layers' functions for edge devices in F-RAN. On the other hand, SDN is based on centralized manner, while F-RAN is based on distrusted manner. Thus, it is still a major challenge that how to achieve SDN ideas in 5G F-RAN. Both of these major research challenges are nontrivial and must be coped for the successful deployment of F-RAN with SDN for 5G systems.

• **Edge Caching:** Edge caching is a key component, which improves the performance of F-RAN by relaxing traffic burden at cloud server, and providing fast content access and retrieval at F-UEs. In F-RAN, the caching space at each F-AP and F-UE is small in comparison to centralized caching mechanism. The edge caching policies, deciding what to cache and when to release cashes in various edge devices are crucial for improving the overall performance of caching in F-RAN. The traditional caching policies in wireless networks, such as first-in first-out, least recently used, and least frequently used, should be evolved to appropriately improve cashe hit ratio in F-RAN.

• **The Virtualization of SDN Controller:** In F-RAN, NFV virtualizes SDN controller in order to run cloud server, which is migrated to fit locations according to the needs of network. However, it is still indistinct due to the distribution characteristic in edge devices that how to virtualize the SDN controller in F-RAN architecture for 5G. Moreover, security, computing performance, portability, VNF interconnection, compatible operation and management with legacy RAN architectures, etc., are still open research challenges.

E. MAJOR RESEARCH CHALLENGES OF TRANSPORT NETWORK IN THE 5G RAN

• **Higher-capacity Backhaul Network:** To carry anticipated huge traffic in future 5G network, it is necessary to employ a backhaul network with enough capacity to connect RAN to the subsequent part of network. According to the literature, 5G backhaul requires ten times higher capacity than legacy LTE network. Massive traffic generated in the 5G RAN brings

many challenges to the backhaul including the network protocols for wireless links, which are needed to be optimized in order to fulfill capacity requirement of massive backhaul traffic.

• **Ultra-low Latency Requirements:** One of the major challenges that remains along with the backhaul is to provide ultra-low latency services for 5G systems. For the time being, mmWave and direct optical fiber are two key enablers for ultra-low latency services, however, it is expected that more options are going to be available in the future. The mmWave faces with propagation challenge and the direct optical fiber is not often cumbersome to lay; therefore, these limitations are required to be solved in order to fulfill ultra-low latency requirement of 5G mobile communication.

• **Elasticity and Flexibility in Backhaul Network:** Future backhaul network requires high flexibility and elasticity. To use the precious resources effectively, joint operation of backhaul and access has to be implemented. To support joint operation, backhaul network has to be flexible in order to ensure the best usage of resources. Backhaul resources should be assigned according to the traffic requirements and unused resources have to be available to be used by other traffic flows, where more resources are required. Hence, proper cooperative, flexible and elastic usage of backhaul resources is unavoidable.

• **Backhaul Syncrhonization:** To benefit from the centralization gains proposed by C-RAN, proper synchronization in the backhaul network is crucial. Literature suggests that LTE level synchronization does not fulfill the requirements of 5G communication systems; therefore, enhanced synchronization is required.

• **Virtualization of Cable Technologies:** In fronthaul, optical fiber is usually preferred to be used due to its high transmission capacity. Meanwhile, other wired transmission media such as coax, twisted pair telephone lines, and power lines can also be deployed as an alternative fronthaul links. So far, there has not been much research attention to the virtualization of cable technologies. In order to fulfill the demand of future heterogeneous mobile networks, we need to change the existing network paradigm from being hardware-centric to software driven. This requires further investigation in terms of both virtualization of network functions and softwarization of cable technologies.

• **Higher-capacity and Lower-latency Fronthaul Links:** The fronthaul link will ask for hundreds of Gbps throughput, which is impractical with current backhaul/fronthaul options. Moreover, more centralization of functions in central unit puts stringent requirement on backhaul latency. Thus, 5G C-RAN will ask for ultra low latency and high data rate in the fronthaul link.

F. MAJOR RESEARCH CHALLENGES OF MMWAVE COMMUNICATION SYSTEM

• **LOS Probability Function:** Obstacles in the environment affect wireless communication channels owing to reflection, diffraction, scattering, absorption, and refraction. From the measurement implemented by New York University, it was observed that reliable communication links can only be established within the range of 200m [201], which is because the millimeter wave signal is sensitive to blockages (buildings, human bodies, etc). Thus, most of the links in mmWave cellular systems are in LOS. The LOS and Non-Line-Of-Sight (NLOS) are illustrated in Fig. 25. The 3GPP standards suggest modeling building blockages by differentiating the LOS and NLOS links using a stochastic model, which is formulated as a probability function of link distance from BS to the typical mobile user in (2). Even though the proposed baseline model is limited to some scenarios with 2-dimensional network models. Most importantly, by plugging the complicated probability function inside SINR model, the tractable analytical model cannot be obtained to evaluate the system performance. Therefore, the most important research challenge in mmWave cellular networks is to propose a suitable LOS probability function in order to capture the property of urban cellular networks and make the obtained performance metrics analytically tractable or mathematically computable:

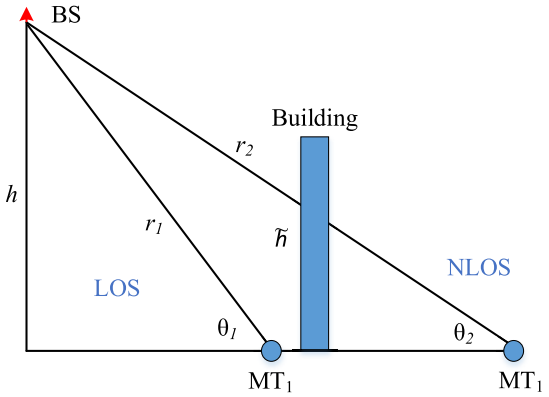

FIGURE 25. Line-of-Sight and Non-Line-of-Sight links.

$$P_{LOS}(r) = \min\left(\frac{18}{r}, 1\right)\left(1 - e^{-\frac{r}{63}}\right) + e^{-\frac{r}{63}} \tag{2}$$

• **Path-loss Model:** Due to the above mentioned blockage effects, where most of the reliable links will be established in LOS, it is necessary to design the small cell networks or ultra-dense networks. The objective of using such deployment is to make mobile users close to the BSs in order to avoid the impact of building blockages. Now with the small cell networks in mmWave system, the conventional unbounded path-loss model will lose its accuracy. Let us explain it as follows. With the conventional power-decaying path-loss model $L(r) = r^{\beta}$, where β is the path-loss slope, the received power is given by

$$P_r = \frac{P_t |h|^2}{r^{\beta}} \tag{3}$$

where $|h|^2$ is the channel fading power. Now in the small cell networks where users are quite close to the BSs, there will be

the following singularity issue:

$$r \to 0 \Rightarrow P_r = \frac{P_t |h|^2}{r^\beta} \Rightarrow \infty \tag{4}$$

This is practically impossible. Thus, the research challenge is to propose a novel path-loss model to avoid such issue and obtain the further insightful guidelines in small cell network designation.

• **Performance Metrics:** In most of the literatures or textbooks, the small-scale fading in wireless channels is assumed to be Rayleigh distributed, which is proved to be invalid for mmWave channels from the measurement results. In mmWave channels, compared with small-scale fading, the impact of shadowing behaves as a significant component in impairing the link reliability due to the large scale building blockage effects. Statistically, the shadowing is modeled as a log-normal distributed random variable with the following Probability Density Function (PDF):

$$f_x(x) = \frac{10}{ln(10)} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(10log_{10}^{x-\mu})^2}{2\sigma^2}\right) \tag{5}$$

where its mean equals to μ and standard deviation equals to σ. The major research challenge by taking into account the impact of shadowing is that how to formulate the performance metrics in a tractable way with the complicated log-normal distribution.

• **Designing of Directional Antenna Array:** Thanks to the small wavelength, mmWave cellular networks are capable of exploiting directional beamforming for compensating of increased path-loss at mmWave frequencies and for overcoming of additional noise due to the large transmission bandwidth. As a desirable bonus, directional beamforming provides interference isolation, which reduces the impact of other-cell interference. However, how to design a directional antenna array to mitigate such effect is still a key research challenge.

G. MAJOR RESEARCH CHALLENGES OF MASSIVE MIMO

• **Reciprocity Calibration:** As discussed in previous section, massive MIMO operates in TDD. The TDD requires reciprocity calibration. There is a need for further investigation to figure out how often the calibration should be done? What is appropriate mechanism to do calibration? And what would be the cost - in terms of time, frequency, resources, and additional hardware components - to do calibration? All these key technical challenges need further comprehensive study.

• **Power Consumption:** One of the advantages of massive MIMO is to offer significantly low radiated power (1000 times) and to simultaneously increase data rates. However, in the real world, total radiated power should be taken into account, including the cost for the processing of baseband signal. In order to reduce internal power consumption, further study is needed for baseband signal processing hardware.

• **Processing of Baseband Data:** The antenna arrays of massive MIMO generate a large amount of baseband data. This data needs to be processed in real time. The processing of the real time baseband data should be as simple as possible - it specifically means that it should be either linear or close to linear. In order to simplify the processing of real time baseband data, further study is needed to be carried out to design optimized algorithms and their deployment in real world.

• **Pilot Contamination:** In massive MIMO systems, every terminal is assigned with an orthogonal uplink pilot sequence. When the pilot sequence of a cell is reused in another cell, the effect of its reusing and other associated with negative consequences are called pilot contamination. The pilot contamination puts a lot of limitations on the deployment of massive MIMO systems in contrast to legacy MIMO systems. Therefore, the pilot contamination needs further investigation in order to find appropriate solutions for existing challenges.

• **Performance Evaluation:** When utilizing massive MIMO instead of conventional MIMO, additional characteristics for channel needed to be considered. Therefore, for realistic performance evaluation of massive MIMO, advanced and sophisticated channel models are required. These models are not necessarily expected to be correct in detailed; however, they should have the fundamental behavior of a channel.

• **Low-cost Hardware:** With the deployment of massive MIMO, the challenge of low-cost hardware also rises. Building tens of Analog-to-Digital (A/D) and Digital-to-Analog (D/A) converters, RF chains, and other components need large amount of CAPEX. Thus, further research is needed to design a system with low CAPEX/OPEX.

H. MAJOR RESEARCH CHALLENGES OF DEVICE TO DEVICE COMMUNICATION

• **Resource Allocation:** In the case of in-band D2D communication, the licensed spectrum allocated to the cellular operators is used by D2D users to communicate with each other. The D2D communication users can reuse uplink/downlink radio resources of the same cell. Therefore, designing of an efficient mechanism of resource allocation for D2D is crucial to avoid interference between D2D users and cellular users. It is worth noting that if D2D communication users are allocated with the resource blocks that are not located nearby to the resource blocks allocated to cellular devices, then, it will significantly reduce interference.

• **Transmission Power of User-devices:** The interference occurs due to inefficient allocation of transmission power. We discussed in previous section that in the case of out-band D2D communication, the unlicensed spectrum adopted by other wireless technologies are used to support direct communication between devices such as WiFi or Bluetooth. Therefore, power allocation in out-band D2D communication seems irrelevant and is not a major concern. However, when it comes to in-band D2D communication, the transmission power of user devices should be allocated properly in order

for the D2D transmitters to not interfere the communication of cellular user devices.

• **Energy Consumption:** One of the main advantages of D2D communication is to improve energy efficiency of user devices. The energy efficiency highly depends on the designing of protocol dedicated for device discovery and D2D communication. For instance, if the protocol is designed in a way to force the user devices to listen for pairing request often or transmit the discovery messages frequently, then, the battery life of user device reduces significantly. Therefore, the trade-off between power consumption of user devices and discovery speed of the user devices needs further studies.

• **System Architecture:** There is a little contribution in terms of designing of a comprehensive architecture to support D2D communication in cellular networks. Therefore, further research seems to be crucial to design a required architecture, which is capable of device discovery, connection setup, interference control, privacy, security, resource allocation, etc., in an efficient way.

• **Application of D2D Communication:** The main purpose of D2D was to assist mobile communication by acting as a relay. However, researchers have proposed various use-cases of D2D in cellular networks such as peer-to-peer communication, multicasting, video dissemination, machine to machine communication and cellular offloading. We believe that D2D concept can be applicable in many other aspects of wireless communication systems such as social networking, vehicular networks, etc. Therefore, further study is required to figure out potential application of D2D communication.

• **Channel Information:** Accurate channel information rooted in efficient interference management, power allocation and radio resource assignment. Traditional cellular networks receive only downlink channel information from user devices, and the uplink channel information is processed at the BS. However, when it comes to D2D communication, we need three types of channel information: the channel gain between D2D pairs, channel between D2D transmitter and cellular user device, and channel gain between cellular transmitter and D2D receiver. The exchange of such a large amount of channel information results in a large overhead to the system, if the system needs instantaneous CSI feedback. The trade-off between accuracy of CSI and its resulting overhead therefore needs further study.

• **Security:** Security and privacy in D2D communications have been widely studied by industry, academia and standardization organizations. However, there are still some concerns, which exist to the application requirements and use-cases of D2D communication. We believe user incentive, requirement gap and conflict, quantification and evaluation tools, and legal and regulation concerns are main aspects of privacy and security of D2D communication that need further studies.

I. MAJOR RESEARCH CHALLENGES OF MASSIVE MACHINE TYPE COMMUNICATION

• **Reducing Random Access Collisions:** The huge amount of devices that share the same radio access can easily overload the Physical Random Access Channel (PRACH) provided by the legacy LTE-A systems; therefore, reliable PHY/MAC solutions to reduce random access collisions are needed.

• **Spectral Efficiency:** Despite of the tiny size of payload transmitted in typical mMTC scenarios, current LTE systems require connected devices to keep exchanging messages with the eNodeB control plane packets that can be significantly larger than the payload, leading to a waste of spectral efficiency.

• **Energy Efficiency:** mMTC use-cases such as IoT and sensor networks deploy massive number of low cost devices that are expected to work over ten years without charging or changing batteries. This asks for solutions of both network access and data transmission with excellent power efficiency.

• **Resource Allocation:** Concerning the coexistence of mMTC with other highly heterogeneous 5G services such as URLLC and eMBB, 5G needs flexible designs of radio resource frames and control channels, in addition to robust waveforms, in order to support a flexible resource allocation among different services.

• **Intra-cell/Inter-cell Interference:** Interference from other devices in both neighbor cells and local cell reduce access rate while increasing power consumption of a user device, especially in case of contentions. Such phenomena emerge in the mMTC scenario as both the number of interference sources and contention risk are high. System-level enhancements of protocol design and power management mechanism are therefore needed to mitigate both intra-cell and inter-cell interference.

J. MAJOR RESEARCH CHALLENGES OF MOBILE EDGE COMPUTING

• **Mobility Management and VM Migration:** Frequent mobility of end-users increases handovers among the edge servers, which affects the QoS. Therefore, efficient mobility management techniques should be designed in order for the end-users to access edge servers seamlessly [285]. Another major challenge is the efficient handling of VM migration procedure when an application is offloaded to several computing nodes. This challenge can be overcome by developing new advanced techniques enabling fast VM migration [265].

• **Business Model:** Edge computing technology introduces a number of potential business opportunities, which requires sustainable business model. Legacy business model cannot fulfill the requirements of edge computing, therefore new business model is expected to support accounting and monitoring for different granularity.

• **Security:** Due to exploitation of mobile device information, the deployment of MEC raises major security concerns to service provider and among end-users. Therefore, MEC requires more stringent security policies since the edge of the network is expected to be handled by the third-party partners.

K. MAJOR RESEARCH CHALLENGES OF NETWORK FUNCTION VIRTUALIZATION

• **Management and Orchestration:** The deployment of NFV in telecommunication network has significant impact on network management systems. It brings major changes in the deployment, operation and management of networks. These changes are required not because of providing of efficient network and services, but also for the sake of exploiting of dynamism and flexibility across the network. It is believed that in such cases, network functions that provide a specific service to a given end-user are going to scattered throughout various server pools in the network. The main challenge is to provide an acceptable level of orchestration for on-demand network functions. Moreover, it also requires efficient management of network functions. Therefore, further studies are required for the management and orchestration of NFV in telecommunication networks.

• **Energy Consumption:** One of the main advantages of NFV deployment in telecommunication networks is to reduce energy consumption. The main reason behind this is the flexibility and ability to scale network resource allocation up and down. Moreover, operators could significantly reduce hardware and physical nodes, therefore, it leads to reduce energy consumption. We believe that energy efficiency of NFV will attract significant attention in telecommunication community. Therefore, further investigations are needed to make sure that there is a balance between performance of function and energy efficiency in the NFV.

• **Resource Allocation:** Before deployment of NFV, the design of physical resource allocation should be considered in order to decrease CAPEX/OPEX, balance load, save energy and recover failures. However, it leads to a new challenge of proposing of efficient algorithms in order to determine which network functions should be placed on which physical resources or which function should be moved from which physical resource to another one. Therefore, to allocate network resources and deploy NFV efficiently, further investigations are required in order to figure out how and based on which mechanisms/algorithms physical resources are needed to be shared among VNFs.

• **Security Threats:** Security is one of the important aspects of NFV supported systems. With the deployment of NFV, new security concerns are emerged such as performance isolation, topology validation and enforcement, availability of management support infrastructure, user authentication, authorization and accounting, etc. Thus, further research is required to provide efficient solutions for existing security threats. The ETSI has established a Security Expert Group (SEG), which is responsible to identify and furthermore resolve existing security threats related to NFV. So far, the SEG has identified many security problems and proposed possible solutions. The SEG has also published NFV security and trust guidance, which plays significant role in the development of NFV, system architecture and operation.

L. MAJOR RESEARCH CHALLENGES OF SOFTWARE DEFINED NETWORKING

• **Deployment:** SDN attracts significant attention to be deployed in different networks such as optical, home, cellular, wireless, etc. Every network has its own specific requirements and characteristics. The deployment of SDN in these networks creates many challenges. The first challenge is that existing network components should have compatibility with SDN-enabled components. The second challenge is to figure out which existing switches or routers should be upgraded. The third challenge is related to logically centralized SDN control. In multiple SDN networks, logically centralized control may not be appropriate, because controls are driven independently by their own controllers. All these three major challenges are caused due to the deployment of SDN in each network, which should be addressed independently for every network in future studies.

• **Data Plane Programmability:** Due to early vision for the SDN-supported system, so far many researches have been carried out to study control plane. However, the data plane programmability also requires investigation. It enriches SDN applications including looking inside the packet, realization in-network caching, compression, and efficient encryption. Without data plane programmability, the flow might be forwarded to the controller, which takes some time and causes inefficiency in the SDN-supported network. Considering the importance of data plane programmability, further researches are required on both data and control planes in parallel in order to deploy innovative applications and services.

• **Performance Evaluation Technique:** SDN implementation in cellular network has raised performance and scalability challenges. So far, few studies have explored performance of SDN-supported systems, however, performance evaluation techniques are mostly wide used ones. Therefore, further work is required to design efficient evaluation techniques in order to avoid long simulation time.

• **API:** The SDN APIs are mainly designed to be used directly by service providers and network operators. However, some APIs are designed to be used by end-users. The APIs dedicated to end-users can be utilized to offer on-demand services. With user-defined controlling, a number of new challenges is raised including the maintenance of baseline fairness and security, elimination the conflict among different end-users, etc.

• **Performance Evaluation Metrics:** Controller platform is considered as a significant component in the SDN-supported system architecture. However, there is still a number of open research challenges including performance, scalability, distribution, modularity, fault-tolerance, load-balance, consistency and synchronization, and highly available programmer-friendly software.

M. MAJOR RESEARCH CHALLENGES OF NETWORK SLICING

• **Service Level Agreement:** The integration of various network slices and partnership between several operators through infrastructure sharing lead to create new challenges for total network investment such as service level agreements between service provider and tenant, expected generating revenue, etc. Considering such an integrated business-oriented approach, new economic strategies and profit modeling should be analyzed and developed in order to meet 5G network requirements. To achieve this goal, a comprehensive study of existing telecom regulatory framework has to be conducted. Moreover, new innovative ways of pricing, cost of infrastructure sharing, service level agreements between the service provider and tenant, and expected generating revenue should be addressed and furthermore standardized.

• **Security Framework:** Existing open interfaces, which support network programmability lead to bring new potential security threats and attacks to softwarized networks. These concerns are raising major barriers on the way to deploy 5G network slicing. Therefore, it calls for an extensive study of multi-level security framework consists of both policies and mechanisms, dynamic threat detection, user authentication, accounting management, and remote attestation.

• **Management and Orchestration:** Despite efficient dynamicity and higher scalability that network slicing brings to 5G system, network management and orchestration in multi-tenant scenarios are still major concerns. In order to dynamically assign network resources to different slices, the optimization policy that manages resource orchestrator should deal with situations where demands are vary. To accomplish this goal, i) an effective cooperation between slice-specific management functional block and resource orchestrator is required, ii) all policies are needed to be automatically validated, and iii) computationally design all resource allocation algorithms and conflict resolution mechanisms at each abstraction layer.

• **Performance Measurement:** When network slices are deployed, network performance analysis and QoS measurement become challenging and complicated tasks. Therefore, an intensive study is required to provide solutions for dynamic performance measurement and network analysis considering both time and cost.

• **Standardization:** The standardization process of network slicing is still at the initial phase. The current status of network slicing standardization is dealing with concept, system architecture, requirements at different network subsections, security, and the impact of slicing on 5G network architecture. Among other standardization organizations, 3GPP is spending significant efforts to develop comprehensive network slice related standards in various working groups such as S1, S2, S3 and S5 in both Rel. 15 and Rel. 16. However, further efforts are required to develop comprehensive global standards of network slicing in 5G communication system.

• **Network Slicing in the RAN:** Network slicing in CN has already been investigated. There is a huge number of research papers that focuses on efficient CN slicing. However, one of the main challenges for further network virtualization lies on the RAN of 5G mobile system. As 5G network is composed of multiple RATs, therefore, it is essential for RAN virtualization solutions to be able to accommodate various 5G use-cases. This presents an additional challenge on the RAN, since it is unclear so far, whether multiple access technologies can be multiplexed over the same hardware or each will need its own dedicated hardware.

XI. CONCLUSIONS

With the dramatically increasing end-user amount, the massive connectivity, the voluminous data, and the new requirements for extremely low latency and higher data rate – 5G claims for a deep rethink on the design of mobile network architecture, more specifically, on its RAN architecture. Motivated by this, we have surveyed existing literature related to RAN architectures for 5G mobile network, namely C-RAN, H-CRAN, V-CRAN and F-RAN. We have also compared them from various perspectives such as energy consumption, operations expenditure, resource allocation, spectrum efficiency, system architecture, and network performance. As a supplement, we have also investigated the key enabling technologies for 5G systems such as MEC, NFV, SDN, and network slicing; and major 5G RATs such as mmWave, massive MIMO, D2D communication, and mMTC. All these technologies play a crucial role in the design of an efficient 5G RAN architecture. Last but not least, we have highlighted some major research challenges in the field of RAN and RATs in 5G systems, and identified several future research directions.

REFERENCES

[1] Cisco. (2017). *Cisco Visual Networking Index: Forecast and Methodology 2016–2021*. Accessed: Sep. 5, 2018. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[2] M. Maternia. (2014). *5G PPP Use Cases and Performance Evaluation Models*. Accessed: Sep. 5, 2018. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf

[3] F. W. Vook, A. Ghosh, and T. A. Thomas, "MIMO and beamforming solutions for 5G technology," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Tampa, FL, USA, Jul. 2014, pp. 1–4.

[4] D. C. Araújo, T. Maksymyuk, A. L. F. de Almeida, T. Maciel, J. C. M. Mota, and M. Jo, "Massive MIMO: Survey and future research topics," *IET Commun.*, vol. 10, no. 15, pp. 1938–1946, Oct. 2016.

[5] J. Rodriguez, "Small cells for 5G mobile networks," in *Fundamentals of 5G Mobile Networks*. Hoboken, NJ, USA: Wiley, 2015, pp. 63–104.

[6] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modelling of base station power consumption under real traffic loads," *Sensors*, vol. 12, no. 4, pp. 4281–4310, Mar. 2012.

[7] 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.

[8] N. Bizanis and F. Kuipers, "SDN and virtualization solutions for the Internet of Things: A survey," *IEEE Access*, vol. 4, pp. 5591–5606, Sep. 2016.

[9] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.

[10] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 84–91, May. 2016.

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

[12] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, G. Wunder, "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.

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

[14] Z. Ma, Z. Zhang, Z. Ding, P. Fan, and H. Li, "Key techniques for 5G wireless communications: Network architecture, physical layer, and MAC layer perspectives," *Sci. China Inf.*, vol. 58, no. 4, pp. 1–20, 2015.

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

[16] N. Panwar, S. Sharma, and A. K. Singh, "A survey on 5G: The next generation of mobile communication," *Phys. Commun.*, vol. 18, pp. 64–84, Mar. 2016.

[17] N. T. Le, "Survey of promising technologies for 5G networks," *Mobile Inf. Syst.*, vol. 2016, Oct. 2016, Art. no. 2676589.

[18] T. K. Sawanobori. (2016). *The Next Generation of Wireless: 5G Leadership in the U.S.* Accessed: Jan. 29, 2018. [Online]. Available: https://www.ctia.org/docs/default-source/default-document-library/5g_white-paper_web2.pdf

[19] European Commission. (2016). *5G for Europe Action Plan*. Accessed: Apr. 11, 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/5g-europe-action-plan

[20] HUAWEI. (2012). *5G: A Technology Vision*. Accessed: Dec. 1, 2018. [Online]. Available: www.huawei.com/ilink/en/download/HW_314849

[21] HUAWEI. (2016). *5G Network Architecture—A High Level View*. Accessed: Dec. 1, 2018. [Online]. Available: http://www.huawei.com/minisite/5g/img/5G_Network_Architecture_A_High-Level_Perspective_en.pdf

[22] 2015. *5G Network Technology Architecture*. Accessed: Dec. 15, 2017. [Online]. Available: www.imt-2020.org.cn/zh/documents/download/8

[23] 5G America. (2017). *5G Spectrum Recommendations*. Accessed: Dec. 26, 2018. [Online]. Available: http://www.5gamericas.org/files/9114/9324/1786/5GA_5G_Spectrum_Recommendations_2017_FINAL.pdf

[24] IMT. (2015). *5G Concept*. Accessed: Dec. 16, 2018. [Online]. Available: www.imt-2020.org.cn/en/documents/download/3

[25] 5G Forum. (2016). *5G Vision, Requirements, and Enabling Technologies*. Accessed: Dec. 18, 2018. [Online]. Available: http://kani.or.kr/5g/whitepaper/5G%20Vision,%20Requirements,%20and%20Enabling%20Technologies.pdf

[26] Ericson. (2017). *5G Systems: Enabling the Transformation of Industry and Society*. Accessed: Jan. 3, 2018. [Online]. Available: https://www.ericsson.com/assets/local/publications/white-papers/wp-5g-systems.pdf

[27] NGMN. (2015). *5G White Paper*. Accessed: Jan. 1, 2018. [Online]. Available: http://www.ngmn.de/fileadmin/ngmn/content/downloads/Technical/2015/NGMN_5G_White_Paper_V1_0.pdf

[28] SK Telecom. (2015). *5G White Paper*. Accessed: Jan. 1, 2018. [Online]. Available: http://www.sktelecom.com/img/pds/press/SKT_5G%20White%20Paper_V1.0_Eng.pdf

[29] University of Surrey. (2016). *5G Whitepaper: The Flat Distributed Cloud (FDC) 5G Architecture Revolution*. Accessed: Dec. 27, 2018. [Online]. Available: https://www.surrey.ac.uk/sites/default/files/5G-Network-Architecture-Whitepaper-(Jan-2016).pdf

[30] 5GMF. (2017). *5G Mobile Communications Systems for 2020 and Beyond*. Accessed: Jan. 3, 2018. [Online]. Available: http://5gmf.jp/wp/wp-content/uploads/2017/10/5GMF-White-Paper-v1_1-All.pdf

[31] Arcip. (2017). *5G: Issues and Challenges*. Accessed: Jan. 1, 2018. [Online]. Available: https://www.arcep.fr/uploads/tx_gspublication/Report-5G-issues-challenges-march2017.pdf

[32] Cisco. (2017). *Cisco Vision: 5G: Thriving Indoors*. Accessed: Jan. 2, 2018. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/5g-ran-indoor.pdf

[33] DoCoMo. (2014). *5G White Paper*. Accessed: Jan. 3, 2018. [Online]. Available: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_5g/DOCOMO_5G_White_Paper.pdf

[34] GSMA. (2014). *Understanding 5G: Perspectives on Future Technological Advancements in Mobile*. Accessed: Jan. 1, 2018. [Online]. Available: https://www.gsmaintelligence.com/research/?file=141208-5g.pdf&download

[35] Qualcomm. (2016). *Whitepaper: 5G-Vision for the Next Generation of Connectivity*. Accessed: Jan. 3, 2018. [Online]. Available: https://www.qualcomm.com/documents/whitepaper-5g-vision-next-generation-connectivity

[36] Samsung. (2015). *5G Vision*. Accessed: Jan. 3, 2018. [Online]. Available: http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf

[37] 5GMF. (2017). *5G in Perspective: A Pragmatic Guide to What's Next*. Accessed: Jan. 3, 2018. [Online]. Available: http://www.skyworksinc.com/downloads/literature/Skyworks-5G%20White-Paper.pdf?source=RF-Cafe

[38] ZTE. (2014). *Driving the Convergence of the Physical and Digital Worlds: White Paper on Next Generation Mobile Technology, the 5G*. Accessed: Jan. 1, 2018. [Online]. Available: http://wwwen.zte.com.cn/en/products/bearer/201402/P020140221415329571322.pdf

[39] Intel. (2012). *5G: A Network Transformation Imperative*. Accessed: Nov. 28, 2018. [Online]. Available: https://www.intel.la/content/www/xl/es/communications/5g-a-network-transformation-imperative.html

[40] G. Auer, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[41] EU Focus Magazine. (2015). *Why the EU is betting big on 5G?* Accessed: Jan. 4, 2018. [Online]. Available: ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=8898

[42] ITIF. (2016). *5G and Next Generation Wireless: Implications for Policy and Competition*. Accessed: Nov. 4, 2018. [Online]. Available: http://www2.itif.org/2016-5g-next-generation.pdf?_ga=2.219840219.397578655.1512407037-1440113886.1512407037

[43] A. Manzalini. (2014). *Software-Defined Networks for Future Networks and Services: Main Technical Challenges and Business Implications*. Accessed: Jan. 3, 2018. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/5g-ran-indoor.pdf

[44] C. Bouras, P. Ntarzanos, and A. Papazois, "Cost modeling for SDN/NFV based mobile 5G networks," in *Proc. 8th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops*, Lisbon, Portugal, 2016, pp. 56–61.

[45] Global Mobile Suppliers Association. (2017). *5G Network Slicing for Vertical Industries*. Accessed: Jan. 19, 2019. [Online]. Available: https://gsacom.com/paper/5g-network-slicing-vertical-industries/

[46] R. Peter, T. Taleb, A. Laghrissi, A. Ksentini, and H. Flinck, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.

[47] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[48] F. Hillebrand, *GSM UMTS: The Creation Global Mobile Communication*. New York, NY, USA: Wiley, 2002.

[49] V. H. M. Donald, "Advanced mobile phone service: The cellular concept," *Bell Syst. Tech. J.*, vol. 58, no. 1, pp. 15–41, Jan. 1979.

[50] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. Hoboken, NJ, USA: Wiley, 2014.

[51] G. Heine and H. Sagkob, *GPRS: Gateway to Third-Generation Mobile Networks*. London, U.K.: Artech House, 2003.

[52] B. Walke, "The roots of GPRS: The first system for mobile packet-based global Internet access," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 12–23, Oct. 2013.

[53] I. Virtej, "Radio resource control for GSM/EDGE Radio Access Network (GERAN)-inter radio access technology and inter-mode procedures," in *Proc. IEEE 54th Veh. Technol. Conf. VTC Fall*, Atlantic City, NJ, USA, vol. 3, Oct. 2001, pp. 1417–1421.

[54] M. Sauter, *From GSM to LTE-Advanced: An Introduction to Mobile Netw. Mobile Broadband*. Hoboken, NJ, USA: Wiley, 2014.

[55] *Technical Specifications and Technical Reports for a UTRAN-based 3GPP System*, document TS 01.01, 3GPP, 2001.

[56] *Technical Specifications and Technical Reports for a GERAN-Based 3GPP System*, document 3GPP TR 41.101, 3GPP, 2006.

[57] *Technical Specifications and Technical Reports for a UTRAN-Based 3GPP System*, 3GPP TR 21.101, 3GPP, 2003.

[58] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. New York, NY, USA: Wiley, 2001.

[59] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE*. New York, NY, USA: Wiley, 2007.

[60] X. Ge, S. Tu, G. Mao, and C. X. Wang, "5G ultra-dense cellular networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[61] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description*, document TS 36.300, 3GPP, Release 8, 2006.

[62] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description*, document TS 36.300, 3GPP, Release 9, 2006.

[63] Alcatel. (2009). *The LTE Network Architecture: A Comprehensive Tutorial*. Accessed: Sep. 19, 2018. [Online]. Available: http://www.cse.unt.edu/~rdantu/FALL_2013_WIRELESS_NETWORKS/LTE_Alcatel_White_Paper.pdf

[64] T. Ali-Yahiya, *Understanding LTE and Its Performance*. Berlin, Germany: Springer, 2011.

[65] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*. New York, NY, USA: Cambridge Univ. Press, 2009.

[66] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description*, document TS 36.300, 3GPP, Release 10, 2009.

[67] 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.

[68] E. Dahlman, *3G Evolution: HSPA LTE for Mobile Broadband*. Amsterdam, The Netherlands: Elsevier, 2010.

[69] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[70] *The Path to 5G Requires a Strong Optical Network: From C-RAN to Cloud-RAN*, Expo, Houston, TX, USA, 2017,

[71] S. Perrin, "Evolving to an Open C-RAN architecture for 5G," Fujitsu Netw. Commun. Inc., Richardson, TX, USA, Tech. Rep., Sep. 2017.

[72] G. Kardaras and C. Lanzani, "Advanced multimode radio for wireless & Mobile broadband communication," in *Proc. Eur. Wireless Technol. Conf.*, Rome, Italy, 2009, pp. 132–135.

[73] *C-RAN the Road Towards Green RAN-White Paper*, China Mobile Res. Inst., China Mobile, Beijing, China, Oct. 2011.

[74] 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.

[75] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[76] V. N. Ha, L. B. Le, and N. Dao, "Energy-efficient coordinated transmission for Cloud-RANs: Algorithm design and trade-off," in *Proc. Annu. Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, 2014, pp. 1–6.

[77] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN," *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 26–32, Jan. 2016.

[78] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Energy-efficient virtual base station formation in optical-access-enabled cloud-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1130–1139, May 2016.

[79] L. Chen, H. Jin, H. Li, J.-B. Seo, Q. Guo, and V. Leung, "An energy efficient implementation of C-RAN in HetNet," in *Proc. IEEE 80th Veh. Technol. Conf.*, Vancouver, BC, USA, Sep. 2014, pp. 1–5.

[80] Z. Tan, C. Yang, and Z. Wang, "Energy evaluation for cloud RAN employing TDM-PON as front-haul based on a new network traffic modeling," *J. Lightw. Technol.*, vol. 35, no. 13, pp. 2669–2677, Jul. 1, 2017.

[81] M. Fiorani, "Modeling energy performance of C-RAN with optical transport in 5G network scenarios," *IEEE OSA J. Opt. Commun. Netw.*, vol. 8, no. 11, pp. B21–B34, Nov. 2016.

[82] "COMP evaluation and enhancement," NGMN, Frankfurt, Germany, White Paper Version 2.0, 2015.

[83] I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: A 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[84] N. Carapellese, "An energy consumption comparison of different mobile backhaul and fronthaul optical access architectures," in *Proc. Eur. Conf. Opt. Commun.*, Cannes, France, 2014, pp. 1–3.

[85] N. Carapellese, M. Tornatore, and A. Pattavina, "Energy-efficient baseband unit placement in a fixed/mobile converged WDM aggregation network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 8, pp. 1542–1551, Aug. 2014.

[86] C. Lim, C. Ranaweera, Y. Yang, and A. Nirmalathas, "Fiber-wireless technology for small cell backhauling," in *Proc. 17th Int. Conf. Transparent Opt. Netw.*, Budapest, Hungary, 2015, pp. 1–4.

[87] F. Tian, P. Zhang, and Z. Yan, "A survey on C-RAN security," *IEEE Access*, vol. 5, pp. 13372–13386, 2017.

[88] B. Niu, Y. Zhou, H. Shah-Mansouri, and V. W. Wong, "A dynamic resource sharing mechanism for cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8325–8338, Dec. 2016.

[89] D. B. Rawat, S. Shetty, and K. Raza, "Secure radio resource management in cloud computing based cognitive radio networks," in *Proc. 41st Int. Conf. Parallel Process. Workshops*, Pittsburgh, PA, USA, 2012, pp. 288–295.

[90] J. You, Z. Zhong, G. Wang, and B. Ai, "Security and reliability performance analysis for cloud radio access networks with channel estimation errors," *IEEE Access*, vol. 2, pp. 1348–1358, 2014.

[91] 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.

[92] L. Chen, L. Liu, X. Fan, J. Li, C. Wang, G. Pan, J. Jakubowicz, and T.-M.-T. Nguyen, "Complementary base station clustering for cost-effective and energy-efficient cloud-RAN," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Computed, Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, San Francisco, CA, USA, Aug. 2017, pp. 1–7.

[93] M. De Andrade, "Cost models for baseband unit (BBU) hotelling: From local to cloud," in *Proc. IEEE 4th Int. Conf. Cloud Netw.*, Niagara Falls, ON, Canada, Oct. 2015, pp. 201–204.

[94] O. Arouk, T. Turletti, N. Nikaein, and K. Obraczka, "Cost optimization of cloud-RAN planning and provisioning for 5G networks," in *Proc. IEEE Int. Conf. Commun.*, Kansas City, MO, USA, May 2018, pp. 1–6.

[95] "Cloud RAN Architecture for 5G-White Paper," Ericsson, Telefonica, Stockholm, Sweden, White Paper, 2017.

[96] H. Niu, C. Li, A. Papathanassiou, and G. Wu, "RAN architecture options and performance for 5G network evolution," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops*, Istanbul, Turkey, Apr. 2014, pp. 294–298.

[97] M. Fathy, B. Mokhtar, M. A. Abdou, and M. R. M. Rizk, "Extended study towards performance improvement of cloud-RAN," in *Proc. 13th Int. Wireless Commun. Mobile Comput. Conf.*, Valencia, Spain, 2017, pp. 1061–1066.

[98] F. A. Khan, H. He, J. Xue, and T. Ratnarajah, "Performance analysis of cloud radio access networks with distributed multiple antenna remote radio heads," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4784–4799, Sep. 2015.

[99] H. Zhang, Y. Dong, J. Cheng, M. J. Hossain, and V. C. M. Leung, "Fronthauling for 5G LTE-U ultra dense cloud small cell networks," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 48–53, Dec. 2016.

[100] U. Karneyenka, "Location and mobility aware resource management for 5G cloud radio access networks," in *Proc. Int. Conf. High Perform. Comput. Simulation*, Genoa, Italy, 2017, pp. 168–175.

[101] D. Naboulsi, "On user mobility in dynamic cloud radio access networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1–9.

[102] L. Liu, F. Yang, R. Wang, Z. Shi, A. Stidwell, and D. Gu, "Analysis of handover performance improvement in cloud-RAN architecture," in *Proc. Int. Conf. Commun. Netw. China*, Kun Ming, China, 2012, pp. 850–855.

[103] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless network virtualization with SDN and C-RAN for 5G networks: Requirements, opportunities, and challenges," *IEEE Access*, vol. 5, pp. 19099–19115, 2017.
[104] *The Benefits of Cloud-RAN Architecture in Mobile Network Expansion*, Fujitsu, Tokyo, Japan, 2015.
[105] "Cloud-RAN: The benefits of virtualization, centralization, and coordination," Ericsson, Stockholm, Sweden, White Paper, 2015.
[106] *Cloud RAN and the Next Generation Mobile Network Architecture*, Huawei, Shenzhen, China, 2017.
[107] *Cloud RAN New Generation Radio Access Network Solution*, ZTE, Shenzhen, China, 2017.
[108] "NFV C-RAN for efficient RAN resource allocation," NEC, Shenzhen, China, White Paper, 2016.
[109] *Towards 5G-RAN Virtualization Enabled by Intel and ASTRI*, Intel, Santa Clara, CA, USA, 2017.
[110] T. Flanagan, "Creating cloud base stations with TI's keystone multicore architecture," Texas Instrum., Dallas, TX, USA, Tech. Rep., Oct. 2011.
[111] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: Next-generation wireless broadband technology [Invited Paper]," *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 10–22, Jun. 2010.
[112] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vijapeyam, T. Yoo, O. Song, D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.
[113] M. Nasimi, F. Hashim, and C. K. Ng, "Characterizing energy efficiency for heterogeneous cellular networks," in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, Pulau Pinang, Malaysia, Dec. 2012, pp. 198–202.
[114] A. Anpalagan, *Design Deployment Small Cell Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[115] H. Claussen, *Small Cell Networks: Deployment, Management, and Optimization*. Hoboken, NJ, USA: Wiley, 2017.
[116] M. Yavuz, F. Meshkati, S. Nanda, N. Johnson, B. Raghothaman, A. Richardson, "Interference management and performance analysis of UMTS/HSPA+ femtocells," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 102–109, Sep. 2009.
[117] H. A. Mahmoud and I. Güvenc, "A comparative study of different deployment modes for femtocell networks," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, Tokyo, Japan, Sep. 2009, pp. 1–5.
[118] C. Patel, M. Yavuz, and S. Nanda, "Femtocells [Industry Perspectives]," *IEEE Wireless Commun.*, vol. 17, no. 5, pp. 6–7, Oct. 2010.
[119] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.
[120] H. Claussen, L. T. W. Ho, and L. G. Samuel, "Financial analysis of a pico-cellular home network deployment," in *Proc. IEEE Int. Conf. Commun.*, Glasgow, Scotland, Jun. 2007, pp. 5604–5609.
[121] H. R. Karimi, "Evolution towards dynamic spectrum sharing in mobile communications," in *Proc. IEEE 17th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Helsinki, Finland, 2006, pp. 1–5.
[122] T. Nihtila and V. Haikola, "HSDPA performance with dual stream MIMO in a combined macro-femto cell network," in *Proc. IEEE 71st Veh. Technol. Conf.*, Taipei, Taiwan, May 2010, pp. 1–5.
[123] 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.
[124] Y. Li, T. Jaing, K. Luo, and S. Mao, "Green heterogeneous cloud radio access networks: Potential techniques, performance trade-offs, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 33–39, Nov. 2017.
[125] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.
[126] R. S. Alhumaima, M. Khan, and H. S. Al-Raweshidy, "Modelling the energy efficiency of heterogeneous cloud radio access networks," in *Proc. Int. Conf. Emerg. Technol.*, Peshawar, Pakistan, 2015, pp. 1–6.
[127] R. S. Alhumaima, M. Khan, and H. S. Al-Raweshidy, "Power model for heterogeneous cloud radio access networks," in *Proc. IEEE Int. Conf. Data Sci. Data Intensive Syst.*, Sydney, NSW, Australia, Dec. 2015, pp. 260–267.
[128] S.-Y. Lien, S.-M. Cheng, K.-C. Cheng, and D. I. Kim, "Resource-optimal licensed-assisted access in heterogeneous cloud radio access networks with heterogeneous carrier communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9915–9930, Dec. 2016.
[129] J. Li, M. Peng, Y. Yu, and Z. Ding, "Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9873–9887, Dec. 2016.
[130] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 879–892, May 2016.
[131] Y. Zhang and Y. Wang, "A framework for energy efficient control in heterogeneous cloud radio access networks," in *Proc. IEEE CIC Int. Conf. Commun. China*, Chengdu, China, Jul. 2016, pp. 1–5.
[132] Q.-T. Vien, T. A. Le, B. Barn, and C. V. Phan, "Optimising energy efficiency of non-orthogonal multiple access for wireless backhaul in heterogeneous cloud radio access network," *IET Commun.*, vol. 10, no. 18, pp. 2516–2524, 2016.
[133] X. He, A. He, Y. Chen, K. K. Chai, and T. Zhang, "Energy efficient resource allocation in heterogeneous cloud radio access networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.
[134] M. Gerasimenko, D. Moltchanov, R. Florea, S. Andreev, Y. Koucheryavy, N. Himayat, S.-P. Yeh, S. Talwar, "Cooperative radio resource management in heterogeneous cloud radio access networks," *IEEE Access*, vol. 3, pp. 397–406, 2015.
[135] L. Yang, "Ergodic rate analysis for user access in downlink heterogeneous cloud radio access networks," in *Proc. IEEE Global Commun. Conf.*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
[136] M. A. Abana, M. Peng, Z. Zhao, and L. A. Olawoyin, "Coverage and rate analysis in heterogeneous cloud radio access networks with device-to-device communication," *IEEE Access*, vol. 4, pp. 2357–2370, 2016.
[137] P. Huang, H. Kao, and W. Liao, "Hierarchical cooperation in heterogeneous cloud radio access networks," in *Proc. IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[138] S. Wang and Y. Sun, "Enhancing performance of heterogeneous cloud radio access networks with efficient user association," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–6.
[139] I. Alqerm and B. Shihada, "Sophisticated online learning scheme for green resource allocation in 5G heterogeneous cloud radio access networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 10, pp. 2423–2437, Oct. 2018.
[140] I. Alqerm and B. Shihada, "Enhanced machine learning scheme for energy efficient resource allocation in 5G heterogeneous cloud radio access networks," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Montreal, QC, Canada, Oct. 2017, pp. 1–7.
[141] Q. Liu, T. Han, N. Ansari, and G. Wu, "On designing energy-efficient heterogeneous cloud radio access networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 721–734, Sep. 2018.
[142] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M.-S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: Advances and challenges," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 66–73, Jun. 2015.
[143] B. Zhang, X. Mao, and J.-L. Yu, "Resource allocation for 5G heterogeneous cloud radio access networks with D2D communication: A matching and coalition approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5883–5894, Jul. 2018.
[144] M. A. Marotta, N. Kaminski, I. Gomez-Miguelez, L. Z. Granville, J. Rochol, L. Da Silva, and C. B. Both, "Resource sharing in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 74–82, Jun. 2015.
[145] 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.
[146] S. Lien, S.-C. Hung, H. Hsu, and K.-C. Chen, "Collaborative radio access of heterogeneous cloud radio access networks and edge computing networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, Kuala Lumpur, Malaysia, May 2016, pp. 193–199.
[147] M. Peng, X. Xie, Q. Hu, J. Zhang, and H. V. Poor, "Contract-based interference coordination in heterogeneous cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1140–1153, Jun. 2015.
[148] M. Peng, H. Xiang, Y. Cheng, S. Yan, and H. V. Poor, "Inter-tier interference suppression in heterogeneous cloud radio access networks," *IEEE Access*, vol. 3, pp. 2441–2455, Dec. 2015.
[149] C. Ran, S. Wang, and C. Wang, "Balancing backhaul load in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 42–48, Jun. 2015.

[150] L. Wang, W. Huang, Y. Fan, and X. Wang, "Priority-based cell selection for mobile equipments in heterogeneous cloud radio access networks," in *Proc. Int. Conf. Connected Vehicles Expo*, Shenzhen, China, 2015, pp. 62–67.

[151] X. Wang, C. Cavdar, L. Wang, M. Tornatore, H. S. Chung, H. H. Lee, S. M. Park, and B. Mukherjee, "Virtualized cloud radio access network for 5G transport," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 202–209, Sep. 2017.

[152] X. Costa-Pèrez, 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.

[153] W. Kiess, P. Weitkemper, and A. Khan, "Base station virtualization for OFDM air interfaces with strict isolation," in *Proc. IEEE Int. Conf. Commun. Workshops*, Budapest, Hungary, Jun. 2013, pp. 756–760.

[154] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Schefczik, and M. Soellner, "Radio base stations in the cloud," *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 129–152, 2013.

[155] A. W. Dawson, M. K. Marina, and F. J. Garcia, "On the benefits of RAN virtualisation in C-RAN based mobile networks," in *Proc. 3rd Eur. Workshop Softw. Defined Netw.*, London, U.K., 2014, pp. 103–108.

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

[157] J. Zeng, X. Su, J. Gong, L. Rong, and J. Wang, "5G virtualized radio access network approach based on NO Stack framework," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–5.

[158] J. Feng, "An approach to 5G wireless network virtualization: Architecture and trial environment," in *Proc. IEEE Wireless Commun. Netw. Conf.*, San Francisco, CA, USA, May 2017, pp. 1–6.

[159] S. Costanzo, "OpeNB: A framework for virtualizing base stations in LTE networks," in *Proc. IEEE Int. Conf. Commun.*, Sydney, NSW, Australia, Jun. 2014, pp. 3148–3153.

[160] H. Li, M. Dong, and K. Ota, "Radio access network virtualization for the social Internet of Things," *IEEE Cloud Computing*, vol. 2, no. 6, pp. 42–50, Nov. 2015.

[161] G. Dandachi, "Joint allocation strategies for radio and processing resources in virtual radio access networks (V-RAN)," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Montreal, QC, USA, Oct. 2017, pp. 1–6.

[162] L. Tian, "Evaluation methodology for virtual base station platforms in radio access networks," *IEEE Access*, vol. 6, pp. 49366–49374, 2018.

[163] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN virtualization in multitenant LTE-A heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6651–6664, Sep. 2016.

[164] I. Al-Samman, "A framework for resources allocation in virtualised C-RAN," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Valencia, Spain, Sep. 2016, pp. 1–7.

[165] D. Mishra, "KORA: A framework for dynamic consolidation & Relocation of control units in virtualized 5G RAN," in *Proc. IEEE Int. Conf. Commun.*, Kansas City, MO, USA, May 2018, pp. 1–7.

[166] W. Al-Zubaedi and H. S. Al-Raweshidy, "A parameterized and optimized BBU pool virtualization power model for C-RAN architecture," in *Proc. IEEE EUROCON 17th Int. Conf. Smart Technol.*, Ohrid, Macedonia, 2017, pp. 38–43.

[167] P. Rost, I. Berberana, A. Maeder, H. Paul, V. Suryaprakash, M. Valenti, D. Wübben, A. Dekorsy, and G. Fettweis, "Benefits and challenges of virtualization in 5G radio access networks," *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 75–82, Dec. 2015.

[168] *Worldwide Internet of Things Forecast Update 2015–2019*, document #US40983216, IDC, Framingham, MA, USA, Feb. 2016.

[169] OpenFog, "OpenFog reference architecture for fog computing," OpenFog Consortium, Fremont, CA, USA, Tech. Rep. OPFRA001.020817, 2017.

[170] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 416–464, 1st Quart., 2018.

[171] M. Mukherjee, L. Shu, and D. Wang, "Survey of fog computing: Fundamental, network applications, and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1826–1857, 3rd Quart., 2018.

[172] R. K. Naha, S. Garg, D. Georgakopoulos, P. P. Jayaraman, L. Gao, Y. Xiang, and R. Ranjan, "Fog computing: Survey of trends, architectures, requirements, and research directions," *IEEE Access*, vol. 6, pp. 47980–48009, 2018.

[173] K. Liang, L. Zhao, X. Zhao, Y. Wang, and S. Ou, "Joint resource allocation and coordinated computation offloading for fog radio access networks," *China Commun.*, vol. 13, no. 2, pp. 131–139, 2016.

[174] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[175] Y.-Y. Shih, W.-H. Chung, A.-C. Pang, T.-C. Chiu, and H.-Y. Wei, "Enabling low-latency applications in fog-radio access networks," *IEEE Netw.*, vol. 31, no. 1, pp. 52–58, Jan. 2017.

[176] S. Jia, Y. Ai, Z. Zhao, M. Peng, and C. Hu, "Hierarchical content caching in fog radio access networks: Ergodic rate and transmit latency," *China Commun.*, vol. 13, no. 12, pp. 1–14, 2016.

[177] S. Kim, "Fog radio access network system control scheme based on the embedded game model," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, pp. 1–14, Dec. 2017.

[178] H. Xiang, "Joint mode selection and resource allocation for downlink fog radio access networks supported D2D," in *Proc. 11th Int. Conf. Heterogeneous Netw. Quality, Rel., Secur. Robustness*, Taipei, Taiwan, Aug. 2015, pp. 177–182.

[179] D. Chen, S. Schedler, and V. Kuehn, "Backhaul traffic balancing and dynamic content-centric clustering for the downlink of fog radio access network," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun.*, Edinburgh, U.K., Jul. 2016, pp. 1–5.

[180] S. Park, O. Simeone, and S. Shamai, "Joint cloud and edge processing for latency minimization in fog radio access networks," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun.*, Edinburgh, U.K., Jul. 2016, pp. 1–5.

[181] S. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, 2016, pp. 315–319.

[182] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[183] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, Jul. 2016, pp. 2029–2033.

[184] S. Yan, M. Peng, M. A. Abana, and W. Wang, "An evolutionary game for user access mode selection in fog radio access networks," *IEEE Access*, vol. 5, pp. 2200–2210, 2017.

[185] X. Wang, S. Leng, and K. Yang, "Social-aware edge caching in fog radio access networks," *IEEE Access*, vol. 5, pp. 8492–8501, 2017.

[186] N. Dao, J. Lee, D.-N. Vu, J. Paek, J. Kim, S. Cho, K.-S. Chung, and C. Keum, "Adaptive resource balancing for serviceability maximization in fog radio access networks," *IEEE Access*, vol. 5, pp. 14548–14559, 2017.

[187] G. M. S. Rahman, M. Peng, K. Zhang, and S. Chen, "Radio resource allocation for achieving ultra-low latency in fog radio access networks," *IEEE Access*, vol. 6, pp. 17442–17454, 2018.

[188] E. Balevi and R. D. Gitlin, "A clustering algorithm that maximizes throughput in 5G heterogeneous F-RAN networks," *Proc. IEEE Int. Conf. Commun.*, Kansas City, MO, USA, May 2018, pp. 1–6.

[189] Y.-J. Ku, D.-Y. Lin, C.-F. Lee, P.-J. Hsieh, H.-Y. Wei, C.-T. Chou, and A.-C. Pang, "5G radio access network design with the fog paradigm: Confluence of communications and computing," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 46–52, Apr. 2017.

[190] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA-based fog radio access networks," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 110–115, Jun. 2018.

[191] H. Zhang, Y. Qiu, X. Chu, K. Long, and V. C. M. Leung, "Fog radio access networks: Mobility management, interference mitigation, and resource optimization," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 120–127, Dec. 2017.

[192] K. Liang, L. Zhao, X. Chu, and H. H. Chen, "An integrated architecture for software defined and virtualized radio access networks with fog computing," *IEEE Netw.*, vol. 31, no. 1, pp. 80–87, Jan. 2017.

[193] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, "Network slicing in fog radio access networks: Issues and challenges," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 110–116, Dec. 2017.

[194] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003–5009, Sep. 2016.

[195] D. Chen, H. Al-Shatri, T. Mahn, A. Klein, and V. Kuehn, "Energy efficient robust F-RAN downlink design for hard and soft fronthauling," in *Proc. IEEE 87th Veh. Technol. Conf.*, Porto, Portugal, Jun. 2018,

[196] Y. Qiu, H. Zhang, K. Long, Y. Huang, X. Song, and V. C. M. Leung, "Energy-efficient power allocation with interference mitigation in MmWave-based fog radio access networks," *IEEE Wireless Commun.*, vol. 25, no. 4, pp. 25–31, Aug. 2018.

[197] T. C. Chiu, W.-H. Chung, A.-C. Pang, Y.-J. Yu, and P.-H. Yen, "Ultra-low latency service provision in 5G fog-radio access networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Valencia, Spain, Sep. 2016, pp. 1–6.

[198] A. S. Marcano and H. L. Christiansen, "Macro cell assisted cell discovery method for 5G mobile networks," in *Proc. IEEE 83rd Veh. Technol. Conf.*, Nanjing, China, May 2016, pp. 1–5.

[199] Z. Qingling and J. Li, "Rain attenuation in millimeter wave ranges," in *Proc. Int. Symp. Antennas, Propag. EM Theory*, Guilin, China, Oct. 2006, pp. 1–4.

[200] R. J. Humpleman and P. A. Watson, "Investigation of attenuation by rainfall at 60 GHz," in *Proc. Inst. Electr. Eng.*, vol. 125, no. 2, pp. 85–91, 1978.

[201] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[202] J. Wang, Z. Lan, C.-W. Pyo, T. Baykas, C.-S. Sum, M. A. Rahman, R. Funada, F. Kojima, I. Lakkis, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[203] Y. Tsang, A. Poon, and S. Addepalli, "Coding the beams: Improving beamforming training in mmwave communication system," in *Proc. IEEE Global Telecommun. Conf.*, Kathmandu, Nepal, 2011, pp. 1–6.

[204] J. Qiao, X. Shen, J. W. Mark, and Y. He, "MAC-layer concurrent beamforming protocol for indoor millimeter-wave networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 327–338, Jan. 2015.

[205] Y. Niu, "A survey of millimeter wave (mmWave) communications for 5G: Opportunities and challenges," in *Wireless Netw.*, vol. 21, no. 8, pp. 1–20, 2015.

[206] D. Wu, J. Wang, Y. Cai, and M. Guizani, "Millimeter-wave multimedia communications: Challenges, methodology, and applications," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 232–238, Jan. 2015.

[207] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, Jan. 2015.

[208] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[209] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[210] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin, and A. Ghosh, "Millimeter wave communications for future mobile networks (guest editorial), part I," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Jul. 2017.

[211] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter wave communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1616–1653, 3rd Quart., 2018.

[212] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, Jr., "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[213] J. Song, L.-T. Tu, and M. Di Renzo, "On the feasibility of interference alignment in ultra-dense millimeter-wave cellular networks" in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2016, pp. 1176–1180.

[214] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[215] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[216] Z. Ding, L. Dai, R. Schober, and H. V. Poor, "NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1879–1882, Aug. 2017.

[217] J. Ma, C. Liang, C. Xu, and L. Ping, "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2696–2707, Dec. 2017.

[218] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 86–94, Jun. 2017.

[219] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for MmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[220] X. Liu, Y. Li, X. Li, L. Xiao, and J. Wang, "Pilot reuse and interference-aided MMSE detection for D2D underlay massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3116–3130, Apr. 2017.

[221] X. Lin, R. W. Heath, and J. G. Andrews, "The interplay between massive MIMO and underlaid D2D networking," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3337–3351, Jun. 2015.

[222] W. Liu, S. Han, C. Yan, and C. Sun, "Massive MIMO or small cell network: Who is more energy efficient?" in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops*, Shanghai, China, Apr. 2013, pp. 24–29.

[223] L. Zhao, H. Zhao, F. Hu, K. Zheng, and J. Zhang, "Energy efficient power allocation algorithm for downlink Massive MIMO with MRT precoding," in *Proc. IEEE VTC-Fall*, Sep. 2013, pp. 1–5.

[224] K. Guo, Y. Guo, G. Fodor, and G. Ascheid, "Uplink power control with MMSE receiver in multi-cell MU-massive-MIMO systems," in *Proc. IEEE ICC*, Jun. 2014, pp. 5184–5190.

[225] H. V. Cheng, E. Bjornson, and E. G. Larsson, "Uplink pilot and data power control for single cell massive MIMO systems with MRC," in *Proc. Int. Symp. Wireless Commun. Syst.*, 2015, pp. 396–400.

[226] T. V. Chien, E. Björnson and E. G. Larsson, "Joint power allocation and user association optimization for Massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6384–6399, Sep. 2016.

[227] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, 4th Quart., 2014.

[228] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2133–2168, 3rd Quart., 2018.

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

[230] M. Noura and R. Nordin, "A survey on interference management for device-to-device (D2D) communication and its challenges in 5G Networks," *J. Netw. Comput. Appl.*, vol. 71, pp. 130–150, Aug. 2016.

[231] F. S. Shaikh and R. Wismüller, "Routing in multi-hop cellular device-to-device (D2D) networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2622–2657, 4th Quart., 2018.

[232] M. Ahmed, Y. Li, M. Waqas, M. Sheraz, D. Jin, and Z. Han, "A survey on socially aware device-to-device communications," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2169–2197, 3rd Quart., 2018.

[233] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in OFDMA cellular networks: A survey and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1885–1922, 4th Quart., 2015.

[234] X. Cheng, Y. Li, B. Ai, X. Yin, and Q. Wang, "Device-to-device channel measurements and models: A survey," *IET Commun.*, vol. 9, no. 3, pp. 312–325, 2015.

[235] O. N. Hamoud, T. Kenaza, and Y. Challal, "Security in device-to-device communications: A survey," *IET Netw.*, vol. 7, no. 1, pp. 14–22, Jan. 2018.

[236] P. Gandotra, R. K. Jha, and S. Jain, "A survey on device-to-device (D2D) communication: Architecture and security issues," *J. Netw. Comput. Appl.*, vol. 78, pp. 9–29, Jan. 2017.

[237] M. Wang and Z. Yan, "A survey on security in D2D communications," *Mobile Netw. Appl.*, vol. 22, no. 2, pp. 195–208, Apr. 2017.

[238] M. Haus, M. Waqas, A. Y. Ding, Y. Li, S. Tarkoma, and J. Ott, "Security and privacy in device-to-device (D2D) communication: A review," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1054–1079, 2nd Quart., 2017.

[239] P. Pirinen, "A brief overview of 5G research activities," in *Proc. Int. Conf. 5G Ubiquitous Connectivity*, Akaslompolo, Finland, 2014, pp. 17–22.

[240] D. Soldani and A. Manzalini, "Horizon 2020 and beyond: On the 5G operating system for a true digital society," *IEEE Veh. Technol. Mag.*, vol. 10, no. 1, pp. 32–42, Mar. 2015.
[241] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
[242] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.
[243] M. T. Islam, A.-E. M. Taha, and S. Akl, "A survey of access management techniques in machine type communications," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 74–81, Apr. 2014.
[244] N. Xia, H. Chen, and C. Yang, "Radio resource management in machine-to-machine communications—A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 791–828, 1st Quart., 2018.
[245] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.
[246] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, A. Ghosh, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.
[247] Z. Dawy, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, 2017.
[248] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.
[249] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, Jan. 2017.
[250] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, C. Zhou, J. Eichinger, and Ö. Bulakci, "On the way to massive access in 5G: Challenges and solutions for massive machine communications," in *Proc. Int. Conf. Cognit. Radio Oriented Wireless Netw.* Cham, Switzerland: Springer, 2015, pp. 708–717.
[251] Y. Meng, C. Jiang, H.-H. Chen, and Y. Ren, "Cooperative device-to-device communications: Social networking perspectives," *IEEE Netw.*, vol. 31, no. 3, pp. 38–44, Mar. 2017.
[252] B. Han and H. D. Schotten, "Grouping-based random access collision control for massive machine-type communication," in *Proc. IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–7.
[253] B. Han, M. A. Habibi, and D. H. Schotten, "Optimal resource dedication in grouped random access for massive machine-type communications," in *Proc. IEEE Conf. Standards Commun. Netw.*, Helsinki, Finland, Sep. 2017, pp. 72–77.
[254] L. Ji, B. Han, M. Liu, and H. D. Schotten, "Applying device-to-device communication to enhance IoT services," *IEEE Commun. Standards Mag.*, vol. 1, no. 2, pp. 85–91, 2017.
[255] Y. Wu, W. Guo, H. Yuan, L. Li, S. Wang, X. Chu, and J. Zhang, "Device-to-device meets LTE-unlicensed," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 154–159, May 2016.
[256] S. Y. Lien, "3GPP device-to-device communications for beyond 4G cellular networks" *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 29–35, May 2016.
[257] V. Sciancalepore, D. Giustiniano, A. Banchs, and A. Hossmann-Picu, "Offloading cellular traffic through opportunistic communications: Analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 122–137, Jan. 2016.
[258] Y. C. Hu, "Mobile edge computing—A key technology towards 5G," ETSI Sophia Antipolis, France, White Paper 11, 2015.
[259] Y. Mao, J. Zhang, Z. Chen, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.
[260] W. Labidi, M. Sarkiss, and M. Kamoun, "Energy-optimal resource scheduling and computation offloading in small cell networks," in *Proc. 22nd Int. Conf. Telecommun.*, 2015, pp. 313–318.
[261] S. Sardellitti, S. Barbarossa, and G. Scutari, "Distributed mobile cloud computing: Joint optimization of radio and computational resources," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 1505–1510.
[262] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.
[263] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading and resource allocation for computation and communication in mobile cloud with computing access point," in *Proc. IEEE Conf. Comput. Commun.*, May 2017, pp. 1–9.
[264] S. Cao, X. Tao, Y. Hou, and Q. Cui, "An energy-optimal offloading algorithm of mobile computing based on HetNets," in *Proc. Int. Conf. Connected Vehicles Expo*, 2015 pp. 254–258.
[265] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.
[266] M. Nasimi, M. A. Habibi, B. Han, and H. D. Schotten, "Edge-assisted congestion control mechanism for 5G network using software-defined networking," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Lisbon, Portugal, 2018, pp. 1–5.
[267] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22–29, Dec. 2016.
[268] S. Vural, P. Navaratnam, N. Wang, C. Wang, L. Dong, and R. Tafazolli, "In-network caching of Internet-of-Things data," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2014, pp. 3185–3190.
[269] L. Gu, "Cost efficient resource management in fog computing supported medical cyber-physical system," *IEEE Trans. Emerg. Topics Comput.*, vol. 5, no. 1, pp. 108–119, Dec. 2017.
[270] G. C. Fox, S. Kamburugamuve, and R. D. Hartman, "Architecture and measured characteristics of a cloud based Internet of Things," in *Proc. Int. Conf. Collaboration Technol. Syst.*, May 2012, pp. 6–12.
[271] *Smart Cells Revolutionize Service Delivery*, Santa Clara, CA, USA, Intel, 2013.
[272] P. A. Frangoudis, L. Yala, A. Ksentini, and T. Taleb, "An architecture for on-demand service deployment over a telco CDN," in *Proc. IEEE Int. Conf. Commun.*, May 2016, pp. 1–6.
[273] S. Retal, M. Bagaa, T. Taleb, and H. Flinck, "Content delivery network slicing: QoE and cost awareness," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–6.
[274] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 59–69, May 2011.
[275] Y. Jararweh, L. Tawalbeh, F. Ababneh, and F. Dosari, "Resource efficient mobile computing using cloudlet infrastructure," in *Proc. IEEE 9th Int. Conf. Mobile Ad-Hoc Sensor Netw.*, Dec. 2013, pp. 373–377.
[276] *Mobile-Edge Computing (MEC); Service Scenarios*, ETSI, Sophia-Antipolis, France, 2015.
[277] X. Luo, "From augmented reality to augmented computing: A look at cloud-mobile convergence," in *Proc. Int. Symp. Ubiquitous Virtual Reality*, Jul. 2009, pp. 29–32.
[278] T. Olsson and M. Salo, "Online user survey on current mobile augmented reality applications," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 75–84.
[279] J. Dolezal, "Performance evaluation of computation offloading from mobile device to the edge of mobile network," in *Proc. IEEE Conf. Standards Commun. Netw.*, Oct. 2016, pp. 1–7.
[280] O. Makinen, "Streaming at the edge: Local service concepts utilizing mobile edge computing," in *Proc. 9th Int. Conf. Next Gener. Mobile Appl., Services Technol.*, Sep. 2015, pp. 1–6.
[281] A. Anjum, T. Abdullah, M. Tariq, Y. Baltaci, and N. Antonopoulos, "Video stream analysis in clouds: An object detection and classification framework for high performance video analytics," *IEEE Trans. Cloud Comput.*, to be published.
[282] R. Yu, Y. Zhang, S. Gjessing, W. Xia, and K. Yang, "Toward cloud-based vehicular networks with efficient resource management," *IEEE Netw.*, vol. 27, no. 5, pp. 48–55, Sep./Oct. 2013.
[283] S. K. Datta, S. Kanti, C. Bonnet, and J. Haerri, "Fog Computing architecture to enable consumer centric Internet of Things services," in *Proc. Int. Symp. Consum. Electron.*, Jun. 2015, pp. 1–2.
[284] S. K. Sharma and X. Wang, "Live data analytics with collaborative edge and cloud processing in wireless IoT networks," *IEEE Access*, vol. 5, pp. 4621–4635, Mar. 2017.
[285] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. 10th Int. Conf. Intell. Syst. Control*, Coimbatore, India, 2016, pp. 1–8.
[286] "Network function virtualisation-introductory white paper" ETSI, SDN and Openflow World Congr., Darmstadt, Germany, 2012.

[287] R. Munoz, R. Vilalta, R. Casellas, R. Martínez, T. Szyrkowiec, A. Autenrieth, V. López, and D. López, "SDN/NFV orchestration for dynamic deployment of virtual SDN controllers as VNF for multi-tenant optical networks," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Los Angeles, CA, USA, 2015, pp. 1–3.
[288] R. Nejabati, S. Peng, M. Channegowda, B. Guo, and D. Simeonidou, "SDN and NFV convergence a technology enabler for abstracting and virtualising hardware and control of optical networks (invited)," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Los Angeles, CA, USA, 2015, pp. 1–3.
[289] I. F. Akyildiz, S.-C. Lin, and P. Wang, "Wireless software-defined networks (wSDNs) and network function virtualization (NFV) for 5G cellular systems: An overview and qualitative evaluation," *Comput. Netw.*, vol. 93, no. 1, pp. 66–79, 2015.
[290] F. Yang, H. Wang, C. Mei, J. Zhang, and M. Wang, "A flexible three clouds 5G mobile network architecture based on NFV & SDN," *China Commun.*, vol. 12, pp. 121–131, Dec. 2015.
[291] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Netw.*, vol. 28, no. 6, pp. 18–26, Nov./Dec. 2014.
[292] J. Matias, J. Garay, N. Toledo, J. Unzilla, and E. Jacob, "Toward an SDN-enabled NFV architecture," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 187–193, Apr. 2015.
[293] W. Ding, W. Qi, J. Wang, and B. Chen, "OpenSCaaS: An open service chain as a service platform toward the integration of SDN and NFV," *IEEE Netw.*, vol. 29, no. 3, pp. 30–35, May 2015.
[294] L. I. B. López, Á. L. V. Caraguay, L. J. G. Villalba, and D. López, "Trends on virtualisation with software defined networking and network function virtualisation," *IET Netw.*, vol. 4, no. 5, pp. 255–263, 2015.
[295] B. Yi, X. Wang, S. K. Das, K. Li, and M. Huang, "A comprehensive survey of network function virtualization," *Comput. Netw.*, vol. 133, pp. 212–262, Mar. 2018.
[296] V. G. Nguyen, A. Brunstrom, K.-J. Grinnemo, and J. Taheri, "SDN/NFV-based mobile packet core network architectures: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1567–1602, 3rd Quart., 2017.
[297] J. G. Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 3, pp. 518–532, Sep. 2016.
[298] Y. Li and M. Chen, "Software-defined network function virtualization: A survey," *IEEE Access*, vol. 3, pp. 2542–2553, 2015.
[299] M. Veeraraghavan, "Network function virtualization: A survey," *IEICE TRANS. COMMUN.*, vol. E100-B, no. 11, pp. 1978–1990, 2017.
[300] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 90–97, Feb. 2015.
[301] I. F. Akyildiz, P. Wang, and S.-C. Lin, "Softair: A software defined networking architecture for 5g wireless systems," *Comput. Netw.*, vol. 85, pp. 1–18, Jul. 2015.
[302] G. Sun, F. Liu, J. Lai, and G. Liu, "Software defined wireless network architecture for the next generation mobile communication: Proposal and initial prototype" *J. Commun.*, vol. 9, no. 12, pp. 946–953, 2014.
[303] D. Kreutz, F. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.
[304] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A survey on software-defined networking,' *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 27–51, 1st Quart., 2014.
[305] R. Masoudi and A. Ghaffari, "Software defined networks: A survey," *J. Netw. Comput. Appl.*, vol. 67, pp. 1–25, May 2016.
[306] H. Farhady, L. HyunYong, and N. Akihiro, "Software-defined networking: A survey," *Comput. Netw.*, vol. 81, pp. 79–95, Apr. 2015.
[307] S. Singh and R. K. Jha, "A survey on software defined networking: Architecture for next generation network," *J. Netw. Syst. Manage.*, vol. 25, no. 2, pp. 321–374, 2017.
[308] Y. Gong, W. Huang, W. Wang, and Y. Lei, "A survey on software defined networking and its applications," *Frontiers Comput. Sci.*, vol. 9, no. 6, pp. 827–845, 2015.
[309] K. Benzekki, A. El Fergougui, and A. E. Elalaoui, "Software-defined networking (SDN): A survey," *Secur. Commun. Netw.*, vol. 9, no. 18, pp. 5803–5833, 2016.
[310] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1617–1634, 3rd Quart., 2014.
[311] M. A. Habibi, "The structure of service level agreement of slice-based 5G network," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, Bologna, Italy, Sep. 2018, pp. 1–6.
[312] M. A. Habibi, B. Han, and H. D. Schotten, "Network slicing in 5G mobile communication: Architecture, profit modeling, and challenges," in *Proc. 14th Int. Symp. Wireless Commun. Syst.*, Bologna, Italy, Sep./Oct. 2017, pp. 1–6.
[313] Q. Li, "An end-to-end network slicing framework for 5G wireless communication systems," 2016, arXiv:1608.00572. Accessed: Aug. 8, 2018. [Online]. Available: https://arxiv.org/abs/1608.00572
[314] O. Sallent, J. Pérez-Romero, R. Ferrús, and R. Agustí, "On radio access network slicing from a radio resource management perspective," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 166–174, Oct. 2017.
[315] I. D. Silva, "Impact of network slicing on 5G radio access networks," in *Proc. Eur. Conf. Netw. Commun.*, Athens, Greece, 2016, pp. 153–157.
[316] Y. L. Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. Int. Conf. Adv. Electr., Electron. Syst. Eng.*, Putrajaya, Malaysia, 2016, pp. 414–420.
[317] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 80–87, May 2017.
[318] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.
[319] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & Prioritization in 5G mobile systems," in *Proc. 22th Eur. Wireless Conf.*, 2016, pp. 1–6.
[320] B. Han, S. Tayade, and H. D. Schotten, "Modeling profit of sliced 5G networks for advanced network resource management and slice implementation," in *Proc. IEEE Symp. Comput. Commun.*, Heraklion, Greece, Jul. 2017, pp. 576–581.
[321] D. Bega, "Optimising 5G infrastructure markets: The business of network slicing," in *Proc. IEEE Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.

MOHAMMAD ASIF HABIBI received the B.Sc. degree in telecommunication engineering from the Information and Communication Technology Institute (ICTI), Kabul University, Afghanistan, in 2011, and the M.Sc. degree in systems engineering and informatics from the Czech University of Life Sciences, Czech Republic, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Wireless Communication, Technische Universität Kaiserslautern, Germany, under the supervision of Prof. H. D. Schotten, where he has been a Marie Curie Research Fellow, since 2017. From 2011 to 2014, he joined Huawei, where he was a Radio Access Network Engineer. His main research interests include software-defined networks, network function virtualization, network slicing, and radio access networks.

MEYSAM NASIMI received the M.Sc. degree in communication engineering from University Putra Malaysia (UPM), in 2014. He is currently pursuing the Ph.D. degree with the Institute of Wireless Communication, Technische Universität Kaiserslautern, Germany, where he has been a Marie Curie Early Stage Researcher, since 2017. His current research interests include wireless communication systems, mobile edge computing, and edge caching.

BIN HAN (M'15) received the B.E. degree in electronic science and technology from Shanghai Jiao Tong University, in 2009, the M.Sc. degree in electrical and information engineering from Technische Universität Darmstadt, in 2012, and the Dr.-Ing. degree in electrical and information engineering from Kalsruher Institut für Technologie, in 2016. He joined the Institute of Wireless Communication, Technische Universität Kaiserslautern, in 2016, and is currently a Senior Lecturer. He has been participating in multiple EU Horizon 2020 research projects for 5G mobile networks. His research interests include communication systems, wireless networks, and digital signal processing, with a current special focus on network slicing and MEC.

HANS D. SCHOTTEN (S'93–M'97) received the Diploma and Ph.D. degrees in electrical engineering from the Aachen University of Technology RWTH, Germany, in 1990 and 1997, respectively. He was a Senior Researcher, the Project Manager, and the Head of the Research Groups, Aachen University of Technology, Ericsson Corporate Research, and Qualcomm Corporate R&D. At Qualcomm, he has also been the Director for Technical Standards and Coordinator of Qualcomm's Activities in European Research Programs. Since 2007, he has been a Full Professor and the Head of the Institute of Wireless Communication, Technische Universität Kaiserslautern. Since 2012, he has been the Scientific Director of the German Research Center for Artificial Intelligence, where he is the Head of the Intelligent Networks Department.