System Architecture and Key Technologies for 5G Heterogeneous Cloud Radio Access Networks

Mugen Peng, Yong Li, Zhongyuan Zhao, and Chonggang Wang

Abstract

Compared with fourth generation cellular systems, fifth generation wireless communication systems are anticipated to provide spectral and energy efficiency growth by a factor of at least 10, and the area throughput growth by a factor of at least 25. To achieve these goals, a H-CRAN is presented in this article as the advanced wireless access network paradigm, where cloud computing is used to fulfill the centralized large-scale cooperative processing for suppressing co-channel interferences. The state-of-the-art research achievements in the areas of system architecture and key technologies for H-CRANs are surveyed. Particularly, Node C as a new communication entity is defined to converge the existing ancestral base stations and act as the base band unit pool to manage all accessed remote radio heads. Also, the software-defined H-CRAN system architecture is presented to be compatible with software-defined networks. The principles, performance gains, and open issues of key technologies, including adaptive large-scale cooperative spatial signal processing, cooperative radio resource management, network function virtualization, and self-organization, are summarized. The major challenges in terms of fronthaul constrained resource allocation optimization and energy harvesting that may affect the promotion of H-CRANs are discussed as well.

With the rapid development of mobile Internet and the Internet of things (IoTs), the demands for high-speed data applications, such as high-quality wireless video streaming, social networking, and machine-to-machine communication, have been growing exponentially recently. It is envisioned that total daily mobile traffic in the representative Western European countries will grow 67 times from 186 terabyte (TB) to 12540 TB from 2010 to 2020. Total worldwide mobile traffic of 351 exabyte (EB) in 2025 will represent a 174 percent increase compared with 2020 [1]. Currently, the cellular networks including the first generation (1G), second generation (2G), third generation (3G), and fourth generation (4G) are far from satisfying the significant traffic increments and the high energy efficiency (EE) because much of the power consumed by a base station (BS) is used to overcome path loss, which in turn causes interferences to other users. The fifth generation (5G) system deployed initially in 2020 is expected to provide approximately 1000 times higher wireless area capacity and save up to 90 percent of energy consumption per service compared with the current 4G system. More than 1000 Gb/s/km^2 area spectral capacity in dense urban environments, 10 times higher battery life of connected devices, and five times reduced end-to-end (E2E) latency are anticipated in 5G systems. The new 5G air interface and spectrum should be combined together with the long term evolution (LTE) and WiFi to provide universal high-rate coverage and seamless user experience [2].

To achieve these goals in 5G systems, advanced radio access technologies and all-Internet Protocol (IP) open Internet network architectures should be evolved smoothly from 4G systems [3]. Accurately, the new breakthroughs in the baseband and radio frequency (RF) are required to enable computationally intensive operations and adapted to new air interfaces in 5G systems. A significant and advanced baseband computation is required to meet the complex requirements of new solutions such as large-scale cooperative signal processing in the physical layer. Meanwhile, the new breakthroughs in the integrated access node and heterogeneous convergence are required to enable the ultra dense radio nodes to work efficiently. The plug-and-play function becomes essential to commercial deployments, in which the available spectral resources should be allocated and the corresponding parameters should be self-organized. Furthermore, the software-defined air interface technologies should be seamlessly integrated into the 5G radio access network (RAN) architectures. The cloud computing based radio access infrastructures would provide on-demand resource processing, delay-aware storage, and high network capacity wherever needed.

Mugen Peng, Yong Li, and Zhongyuan Zhao are with Beijing University of Posts and Telecommunications.

Chonggang Wang is with InterDigital Communications.

5G C-RAN Solution

Some advanced technologies, such as the cloud radio access network (C-RAN) and ultra small cells based heterogeneous network (HetNet), have been presented as potential 5G solutions. C-RAN has attracted intense research interest from both academia and industry (such as China Mobile, Huawei, Alcatel Lucent, Qualcomm) [4]. In C-RANs, a large number of low-cost remote radio heads (RRHs) are randomly deployed and connect to the base band unit (BBU) pool through the fronthaul links. C-RANs have several advantages. First, by moving RRHs closer to the users, higher system capacity and lower power consumption can be achieved because the signal does not need to propagate a long distance to reach the users. Second, since the baseband processing is centralized at the BBU pool, the cooperative processing techniques to mitigate interferences can be leveraged. Third, by exploiting the resource pooling and statistical multiplexing gain, C-RAN is much more efficient in both energy and cost aspects because it is not necessary to dimension the computing resources of each traditional BS according to the individual peak load. However, the fronthaul constraints have great impact on worsening performances of C-RAN, and the scale size of RRHs accessing the same BBU pool is limited and could not be too large due to the implementation complexity. Furthermore, many kinds of system architectures have been proposed by different mobile operators, manufacturers, and even research institutions to explore potential advantages of C-RANs. Therefore, a unified C-RAN for 5G is still not straightforward.

5G HetNet Solution

To increase the capacity of cellular networks in dense areas with high traffic demands, the low power node (LPN) serving for the pure "data-only" service with high capacity is identified as one of the key components in HetNets [5]. One key advantage of HetNets is to decouple the control plane and user plane. LPNs only have the control plane, while the control channel overhead and cell-specific reference signals of LPNs can be fully shifted to macro base stations (MBSs). Unfortunately, an underlaid structure in which MBSs and LPNs reuse the same spectral resources could lead to severe inter-tier interference. Hence, it is critical to suppress interference through advanced signal processing techniques to fully unleash the potential gains of HetNets, such as adopting the advanced coordinated multi-point (CoMP) transmission and reception technique to suppress both intra-tier and inter-tier interference. It was reported that the average spectral efficiency (SE) performance gain from the uplink CoMP in downtown Dresden field trials was only about 20 percent with non-ideal backhaul in [6].

The 5G H-CRAN Solution

To fulfill new breakthroughs anticipated in 5G systems, and to overcome the aforementioned challenges in both C-RANs and HetNets, we presented heterogeneous cloud radio access networks (H-CRANs) as 5G RANs in our prior work [7], which are fully backward compatible with different kinds of C-RANs and HetNets. The motivation of H-CRANs is to embed the cloud computing technology into HetNets to realize the large-scale cooperative signal processing and networking functionalities, and thus SE and EE performance is substantially improved beyond existing HetNets and C-RANs. In the H-CRAN based 5G system, the control plane and user plane are decoupled. MBSs are mainly used to deliver control signaling of the whole H-CRAN and provide seamless coverage, while RRHs are used to provide high speed data transmission in the hot spots with huge data services. The large-scale cooperative signal processing and networking techniques in H-CRANs are threefold. The first comprises advanced spatial signal processing techniques in the physical layer (PHY), including the centralized massive multiple-input-multiple-output (MIMO) and the distributed large-scale spatial cooperative processing. The second comprises large-scale cooperative radio resource management and cloudization in the medium access control (MAC) and upper layers. The third comprises intelligent and self-organizing network management in the network layer to support self-configuration, self-optimization, and self-healing in the ultra dense communication scenario with a huge number of nodes deployed randomly.

Motivation and Organization

In this article we are motivated to make an effort to offer a comprehensive discussion on recent advances of system architectures and technological principles in 5G H-CRANs. Specifically, the application scenarios and the H-CRAN system architecture are presented, where the new communication entity Node C is defined and the software-defined H-CRAN architecture is designed. The large-scale cooperative signal processing and networking techniques strictly related to improve SE and EE in H-CRANs are surveyed, including advanced spatial signal processing, cooperative radio resource management, network function virtualization (NFV), and self-organizing network (SON). The challenging issues related to resource allocation optimization and energy harvesting are discussed as well.

The remainder of this article is organized as follows. The application architecture and system components of H-CRANs are introduced in the next section. Then the promising large-scale cooperative signal processing and networking technologies are discussed. Future challenges will be highlighted after that, and our conclusions are presented in the final section.

Application Architecture and System Components

H-CRAN based 5G systems have capabilities to provide anytime, anywhere gigabit data rate service to the desired user equipment (UE), where an "edgeless" experience to UEs is provided under the help of heterogeneous MBSs and ultra dense RRHs. Based on cloud computing technologies, the on-demand resource processing, storage, and network capacity wherever needed are fulfilled. Software-defined air interfaces and networking technologies are seamlessly integrated into the H-CRAN architectures, which enables the flexibility to create new services and applications.

Application Architecture of H-CRANs

As shown in Fig. 1, H-CRANs take full advantages of cloud computing and heterogeneous convergence technologies, where the new communication entity Node C (Node with cloud computing) as the evolution of BS in the 3rd generation project partnership (3GPP) is presented to converge different RANs for the existing legacy communication entities (ACEs, i.e. MBSs, micro BSs, pico BSs, etc.) and offers processing and networking functionalities in the physical and upper layers for the newly designed RRHs. When Node C works to converge ACEs, it can be regarded as a convergence gateway to execute the cooperative multiple-radio resource managements (CM-RRM) and media independent handover (MIH) functionalities, and the functionalities of a traditional radio network controller (RNC) and BS controller (BSC) can be embodied into Node C. When Node C is used to manage RRHs, it acts as the BBU pool, which is inherited from C-
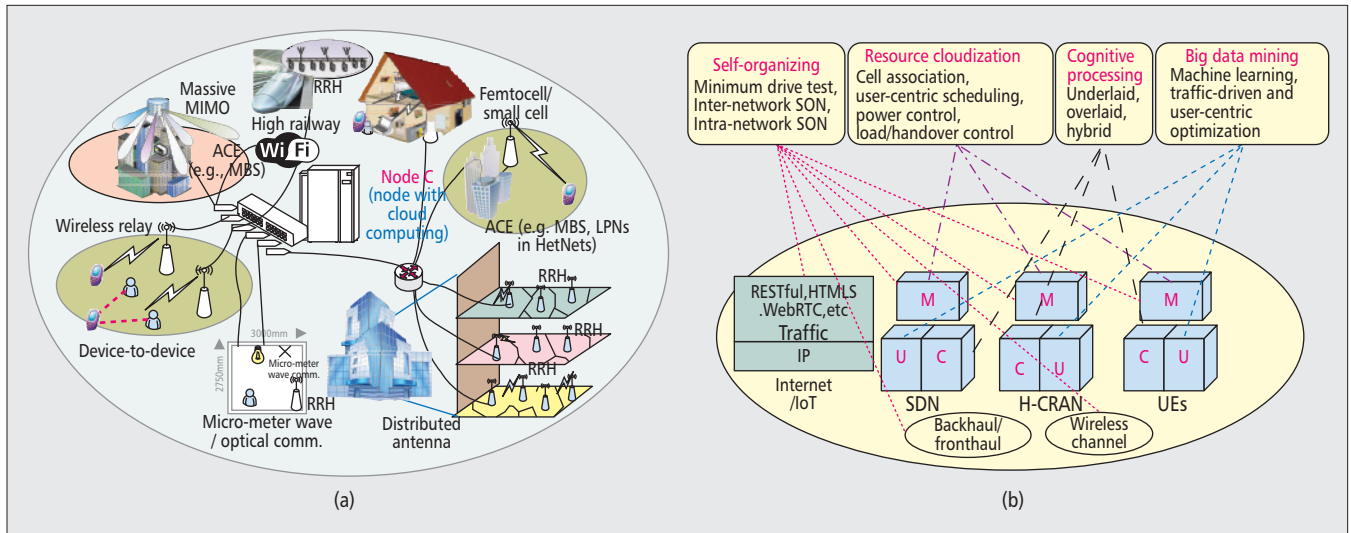
Figure 1. Application architecture, system components, and key functionalities in H-CRANs. a) Application scenarios of 5G systems; b) 5G components and key functionalities.

RANs. In addition, Node C has powerful computing capabilities to execute large-scale cooperative signal processing in the physical layer and large-scale cooperative networking in the upper layers.

In H-CRANs, RRHs are mainly used to provide high speed data transmission without the control plane in hot spots. The control channel overhead and cell specific reference signals for the whole H-CRAN are delivered by ACEs. To guarantee seamless coverage, UEs nearer to ACEs than RRHs are served by ACEs and called HUEs. To improve the intelligence and self-organizing capabilities, the universal plug-and-play functionality is critical, which allows both the newly deployed RRHs and the existing ACEs to access the Node C instantaneously and automatically. With the help of Node C in H-CRANs, hundreds of RRHs and several tens of ACEs can be served simultaneously, and the control plane and user plane separation can be realized efficiently.

In view of the problems existing in Internet Protocol (IP) networks, software-defined network (SDN) is proposed as the IP evolution in 5G systems to simplify the network construction and operational costs by adopting NFV technology [8]. The SDN based wireless network infrastructures are constantly emerging, in which SDN is co-located with the Internet/IoT network entities and decentralizes RRHs/ACEs closer to the desired UEs. To be compatible with the SDN architecture, the adaptive signaling/control mechanism between connection-oriented and connectionless is supported in H-CRANs, which can achieve significant overhead savings in the radio connection/release by moving away from a pure connection-oriented mechanism. Node C is interfaced to SDN to complete NFV functionalities, which is urgent to be adaptive to the dramatically growing traffic features and demands. Different types of radio communication technologies in the physical layer can be adopted in RRHs to improve transmission bit rates, such as IEEE 802.11 ac/ad, millimeter wave communication, and even optical light communication.

To further decrease the power consumption in H-CRANs, the activating RRHs are adaptive to the traffic volume. When the traffic load is low, some potential RRHs fall into sleep mode under the supervision of Node C. However, when the traffic load becomes tremendous in a small special zone, both the heterogeneous ACEs and dense RRHs work together to satisfy the huge capacity requirements. Further, based on NFV technology, RRHs with high load can borrow the radio resources from neighboring homogeneous and heterogeneous RRHs with low load.

System Components of H-CRANs

As shown in Fig. 1b, the H-CRAN based 5G system is comprised of UEs, H-CRAN, SDN, and Internet/IoTs. There exists three planes in this architecture: the user (U) plane, the control (C) plane, and the management (M) plane. Each of them carries out different areas of operations and processes different types of functionalities. Generally speaking, the U plane carries the actual user traffic and executes the related traffic processing to satisfy various quality of service (QoS) requirements. The C plane carries the control signaling, and takes charge of resource allocation and traffic processing to improve SE and EE. The M plane executes administration and operation, and is mainly responsible for adding, deleting, updating, and modifying the logic and interactions for the U plane and the C plane. Except for Internet/IoT networks, the other three components, i.e. UE, H-CRAN, and SDN, can be vertically divided into U, C and M planes. It is worth noting that only the U plane is configured in RRHs, while the C plane and M plane are incorporated in ACEs.

The H-CRAN architecture is software defined, and is characterized by attributes of both SDN and cloud computing. The network control information is delivered from SDN to Node C through the standardized southbound interface, i.e. OpenFlow [8]. By centralizing the network intelligence, the decision-making is facilitated from a global viewpoint. These features are in contrast with current wireless networks, which are built on an autonomous view, and where nodes are unaware of the overall state of the network. Meanwhile, the presented architecture is inherently controlled by the software defined functionality. The centralized programmable architecture enables the network configuration to be automated under the rapid adoption of cloud computing. By providing open application programming interfaces (APIs) for Internet/IoT applications to interact with the other entities in 5G systems, H-CRAN can achieve unprecedented innovation and differentiation. This software-defined H-CRAN architecture brings significant benefits compared with traditional architectures. On the one hand, the programmability of the logically centralized controller makes it convenient to incorporate new ideas in the network and offers network managers the flexibility to configure, manage, and optimize network resources via dynamic, self-organizing programs. On the other hand, benefitting from the centralized approach, operators can make network-wide traffic forwarding decisions in the logically centralized controller instead of configuring all network
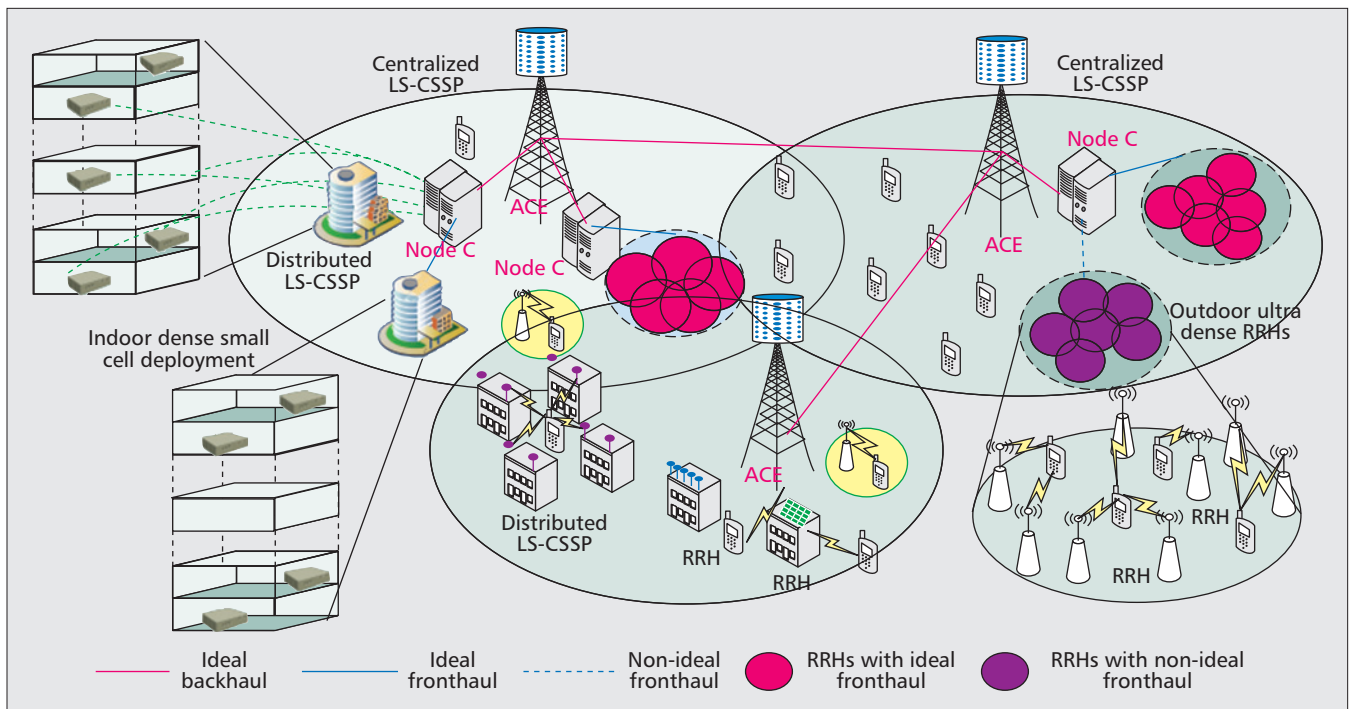
Figure 2. Application architecture, system components, and key functionalities in H-CRANs.

devices individually to make changes in network behavior.

To improve SE and EE performance of the software-defined H-CRAN, there are four key functionalities to be implemented, i.e. self-organizing, radio resource cloudization, cognitive processing, and big data mining. The big data mining functionality imports machine learning technology into cooperative signal processing and intelligent networking, which makes the 5G system adaptive to the Internet/IoT traffic's features and demands. The H-CRAN based 5G systems are traffic-driven and user-centric. To avoid inter-tier interference, the cognitive processing technique is utilized to make RRHs and ACEs work cooperatively in the overlaid scenario when the overall system load is not high. Furthermore, to improve SE performance when the overall system load is high, the radio resource cloudization technique is used to decrease the inter-tier interference and improve the flexibility of radio resource's reuse. To decrease the operation cost and enhance the networking intelligence, the self-organizing H-CRAN is necessary. With measurement and analysis of SE and EE performance, self-organizing functionality automatically configures and optimizes the traffic, fronthaul and radio resources, and regulates operations of the M plane without human intervention.

Promising Key Technologies

To take full advantage of H-CRANs, large-scale cooperative spatial signal processing (LS-CSSP) and cooperative networking techniques should be exploited. The LS-CSSP technique is utilized to fulfill the interference cancelations through collaboration over the software-defined architectures. User-centric and delay-aware large-scale cooperative radio resource management (LS-CRRM) should be highlighted. Further, to be compatible with the SDN based 5G core networks, heterogeneous radio resources should be reused and cloudized, and NFV should be exploited and realized. Considering the complexity and cost of network planning and maintenance, the large-scale self-organizing network (LS-SON) is indispensable to improve intelligence and lower costs. Some other technologies impacting the performance of H-CRANs that are not

introduced in this Section, such as fronthaul compression and quantization, dynamical cell clustering, and channel estimation, are still challenging topics for research in the future.

Large-Scale Cooperative Spatial Signal Processing (LS-CSSP)

For the ultra dense nodes deployed in the traditional cellular networks or HetNets, although the impact of large-scale fading is alleviated, the existence of severe co-channel interferences becomes the performance bottleneck. The LS-CSSP, including both distributed and centralized modes, is prominent to improve SE performance with the gigabit rate as shown in Fig. 2. The distributed LS-CSSP relies on spatial multiplexing, which in turn shares the same radio resources among RRHs. By transmitting collaboratively among multiple RRHs, which can be treated as a virtual antenna array, the interference can be avoided spatially without occupying extra radio resources.

For the distributed LS-CSSP, it is necessary to exploit cooperative precoding schemes that make special considerations of capacity constraints on fronthaul, implementation complexity, and difficulty in obtaining perfect channel state information (CSI) for all connected RRHs and ACEs. Though dirty paper pre-coding can achieve a good capacity region, it is impractical because complexity increases with the number of connected RRHs and ACEs. Linear precoding is a good alternative with less complexity, which can be optimized jointly with the combination of constraints on fronthaul [9]. Meanwhile, cell association has great impact on precoding performance, and the joint cell selection and precoding is key to improve EE and SE performance. As shown in Fig. 3 [10], in the uplink, the distributed LS-CSSP has good potential to strengthen uplink signals and suppress inter-RRH interference, whose gain can reach 40 to 100 percent in the weak coverage area (that is the area where the reference signal receive power (RSRP) is lower than –95dBm), and in some areas the gain can exceed 100 percent. Due to the severe inter-RRH interference, the uplink throughput cannot reach the peak at a good coverage area (e.g. where RSRP is higher than
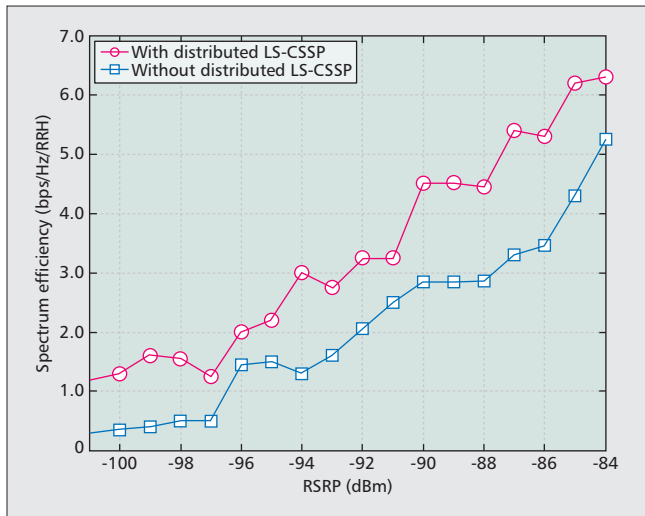
Figure 3. Uplink Throughput Gains from distributed LS-CSSP, where the inter-site distance is around 1 kilometer, two cells are dynamically chosen to perform the uplink LS-CSSP in the collaborative cluster with six cells, and five interfering users in the surrounding cells with 50 percent resource block (RB) utilization ratio are configured3 [10].

–90dBm) when the distributed LS-CSSP is not utilized. Distributed LS-CSSP can bring 1 to 2 Mb/s throughput boost and result in 20 to 50 percent throughput gain at the test area.

The superior centralized processing capability of H-CRAN provides great convenience for making distributed LS-CSSP effective, and can even support the combination with other advanced techniques such as interference alignment. However, some existing technical issues are still challenging. One of the key challenging problems is to determine the scale size of distributed LS-CSSP adaptively. Due to the requirement of a large amount of CSIs, there exists a trade-off between the SE gain and scale size of the distributed LS-CSSP. On the one hand, the improvement of SE performance might be limited when the scale of the distributed LS-CSSP is small. On the other hand, by allowing the participants overly, the signaling exchanging is intractable, and the accuracy and instantaneity of CSI decline as well, which degrades SE performance severely. Therefore, it is critical to clarify the boundary conditions in which the distributed LS-CSSP can achieve significant gains in H-CRANs.

By enlarging the special degrees of freedom, MIMO can improve SE and EE performance by achieving an extra diversity-multiplexing trade-off. The centralized LS-CSSP is equipped with hundreds of low-power antennas at a co-located ACE site, which is mainly used to improve capacity, extend coverage, and decrease antenna deployment complexity. According to the law of large numbers, the channel propagation condition can be hardened, which ensures that the transmission capacity increases linearly as the number of antennas increases. The centralized LS-CSSP can increase the capacity 10 times or more, and improve the radiated EE performance on the order of 100 times as well [11].

Unlike RRHs, ACEs do not need to upload all their observations to Node C for the joint signal cooperative processing. By deploying ACEs with the centralized LS-CSSP, instead of using a huge number of RRHs in some coverage areas, the overload on fronthauls between RRHs and Node C is released, and the time latency of backhaul between MBSs and Node C is alleviated. Compared with ACEs without the centralized LS-CSSP, or the conventional C-RAN configuration, ACEs with the centralized LS-CSSP reduce inter-tier interference to adjacent RRHs/ACEs because they can serve a large area with diluting the density of active RRHs. Moreover, the cooperative beamforming between ACEs and RRHs can suppress remaining inter-tier interference.

Large-Scale Cooperative Radio Resource Management (LS-CRRM)

To optimize the performance of H-CRANs, more advanced radio resource management schemes are required than the traditional cellular network due to the additional computing resources. Since the physical radio resources can be completely reused among RRHs, inter-tier interference between RRHs and ACEs should be suppressed with advanced LS-CRRM techniques. Traditional soft fractional frequency reuse (S-FFR) is considered to be an efficient inter-cell and inter-tier interference coordination technique, in which the service area is partitioned into spatial regions, and each subregion is assigned different frequency sub-bands. Therefore, the cell-edge-zone UEs do not interfere with cell-center-zone UEs, and with an efficient channel allocation method, the cell-edge-zone UEs may not interfere with neighboring cell-edge-zone UEs [12]. Considering that the ACE is mainly used to deliver control signaling and guarantee seamless coverage, the QoS requirement of HUEs is not high, while the QoS requirement for UEs accessing RRHs (denoted as RUEs) is with a high priority. Consequently, an enhanced S-FFR scheme should be proposed to suppress the inter-tier interference between the ACE and RRHs in Fig. 4a, where only partial radio resources are allocated to both RUEs and HUEs with the low QoS requirement, and other radio resources are allocated to RUEs with high QoS requirements.

For the enhanced S-FFR, the RUEs with low QoS requirements will share the same radio resources with HUEs, which is absolutely different from that in the traditional S-FFR, where the radio resources allocated to cell-edge-zone UEs are orthogonal. If the traditional S-FFR is utilized in H-CRANs, only the cell-center-zone RUEs will share the same radio resources with HUEs, which decreases SE performance significantly. Further, how to determine UEs located in the cell-edge or cell-center zone is challenging for the traditional S-FFR. Fortunately, these problems are solved via enhanced S-FFR, where only the QoS requirement should be distinguished for RUEs. To avoid inter-tier interference between RRHs and ACEs, the outband frequency is preferred to be used according to the standards of HetNets in 3GPP, which means that RBs for ACEs should be different from those for RRHs. Meanwhile, to save the occupied frequency bands, the inband strategy is also defined in 3GPP, which means that both RUEs and HUEs will share the same RBs even when the inter-tier interference exists. To be completely compatible with inband and outband strategies in 3GPP, only two RB sets (i.e. Ω_1 and Ω_2) should be divided for the enhanced S-FFR scheme. Obviously, if the locations of RUEs could be known and the traffic volumes in different zones are clearly anticipated, more RB sets can be divided in Ω_1 to achieve a higher performance gain. Ω_1 is only allocated to RUEs with high rate-constrained QoS requirements, and Ω_2 is allocated to RUEs and HUEs with low rate-constrained QoS requirements. Since all signal processing for different RRHs is executed on the BBU pool centrally, inter-RRH interference can be avoided and the same radio resource can be shared among RRHs, thus the radio resources among RRHs are reused and virtualized. Based on the enhanced S-FFR, inter-tier interference among RRHs and ACEs can be suppressed efficiently.

To avoid traffic congestion and obtain the near-optimal performance in H-CRANs, the LS-CRRM scheme should take the queue state information (QSI) into account to guar-
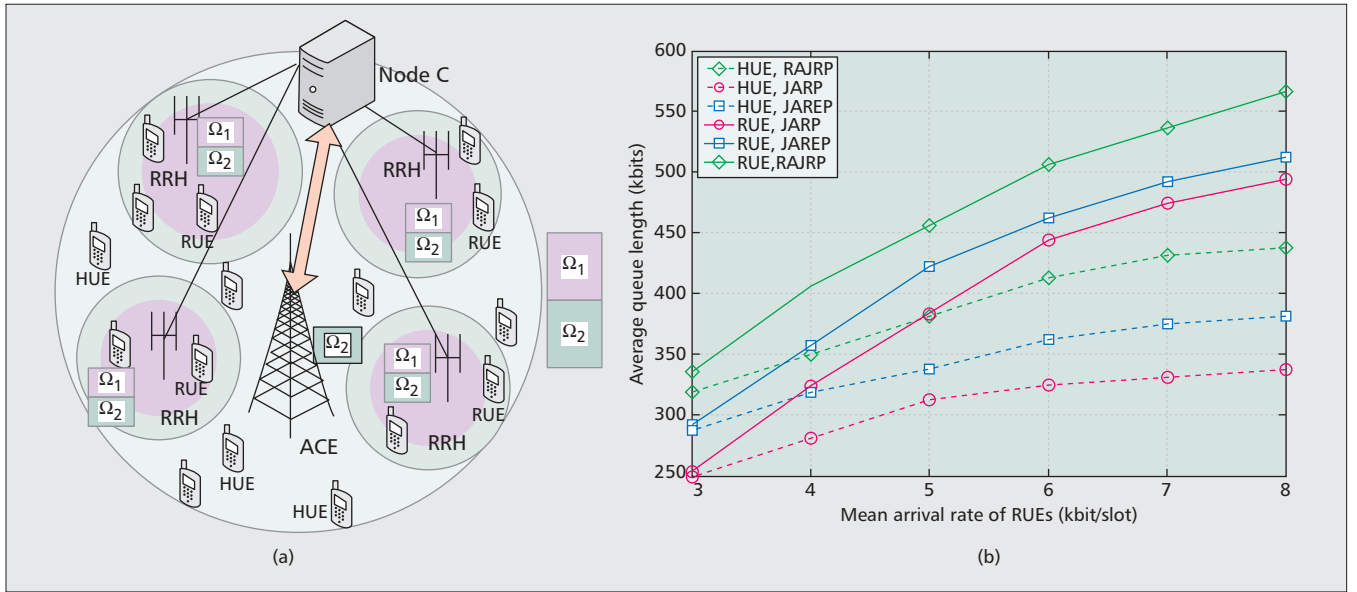
Figure 4. Enhanced soft FFR and performance evaluations of LS-CRRM schemes in H-CRANs, where H-CRAN consists of 1 ACE and 4 RRHs, the number of RUEs is 4, the number of HUEs is 8, the number of RBs is 10, the bandwidth of each RB is 15kHz, and the mean traffic arrival rate of HUEs as 3 kb/slot [13]: a) Enhanced soft fractional frequency reuse; b) Average queue length vs mean arrival rate of RUEs.

antee fairness among UEs and improve the utilization efficiency of radio resources. Particularly, the joint optimization problem taking traffic admission control, RRH/MBS association, RB, and power allocation into account subject to the average and instantaneous power consumption constraints should be researched, and performance of the corresponding joint-association-RB-power allocation (JARP) scheme are evaluated in Fig. 4b, where the random-association-joint-RB-and-power allocation (RAJRP) scheme and joint-association-and-RB-equal-power allocation (JAREP) scheme are compared as baselines [13]. For RAJRP, the HUE randomly accesses the target RRH with the probability proportional to the throughput virtually, and the RB and power are jointly optimized using our algorithm. For JAREP, association and RBs are jointly optimized using our algorithm, and the maximum average power is equally allocated on all RBs of each RRH.

The performance of average queue length of HUEs and RUEs versus mean traffic arrival rate of RUEs are compared among these three schemes. It is observed that the bigger the traffic arrival rate of RUEs, the larger the average queue length for HUEs. This can be understood by the fact that more RBs and power are allocated to RUEs to maintain queue stability, leaving fewer resources for HUEs that are intended to access RRHs. As the mean arrival rate of RUEs increases, the system throughput utility increases, but with a diminishing slope. The rationale is that all traffic arrivals of RUEs can be admitted by these schemes according to the threshold-based admission control policy when the traffic arrival rate of RUEs is relatively small, and some traffic arrivals for RUEs have to be denied to stabilize the traffic queues with higher traffic arrival rate of RUEs. JARP achieves a higher utility and smaller average queue length than JAREP and RAJRP under all traffic arrival rates, which implies the importance of the joint optimization of association, RB, and power allocation.

Network Function Virtualization (NFV) in H-CRANs

Based on the software-defined H-CRAN architecture, NFV is an emerging technology for cost-effective sharing of scarce radio and computing resources in 5G systems, which aims to evolve standard computing virtualization technology to con-solidate network equipment into high volume servers, switch-es, and storage. Accurately, as highly complementary to the control and user plane separation in H-CRANs, NFV in SDNs is applicable to the packet processing in the U plane and the control function in the C plane, which offers the potential capabilities for both enhancing service delivery and reducing overall costs. By enabling NFV with OpenFlow-enabled SDN, network operators can realize significant bene-fits through utilizing cloud computing technology.

H-CRAN infrastructures provide potential capabilities to enhance resource availability and usage by means of orchestration and management mechanisms, which are applicable to the automatic instantiation of virtual appliances and the management of resources by assigning virtual appliances to the correct computing core, memory, and interfaces. The NFV designed for H-CRANs is shown in Fig. 5, where the physical radio resources are cloudized, and computing resources in both intra-RANs and inter-RANs are virtualized. The functionalities of the virtual machine (VM) technology executed in Node C include:

• Allow for quick creation and deletion of VM instances for management purposes.
• Support a wide variety of distributions to allow customization of the environment.
• Allow the user to customize everything from the kernel to the drivers.
• Provide the means for easy administration.

To fulfill the virtual infrastructure, the Node C should embody the generic high-volume computing servers, storage devices, and high-volume network switches, and be organized by the orchestration.

An important aspect for the NFV is the transparency, i.e. the virtual nodes cannot see or exchange any type of information, in order to assure isolation of the heterogeneous convergence. By using the virtual function in Fig. 5, the heterogeneous of ACEs and RRHs are virtualized and are transparent to the lower layers. Additionally, the data exchange in the virtual layer is transparent to the physical provider to preserve the privacy of customers. Nevertheless, some minimal primitives to inspect the activity of different slices are normally available. As an example, the controller of physical radio
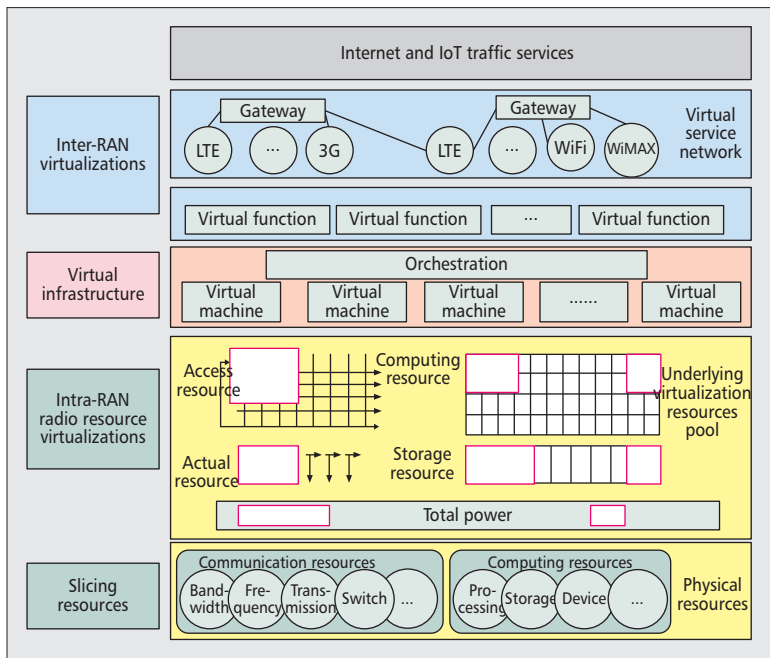
Figure 5. Network function virtualization (NFV) in H-CRANs.

resources is allowed to know the actual usage of computational resources and traffic consumption.

The design and implementation of NFV in 5G systems should highlight the coexistence of slices with bandwidth-based and resource-based reservations because the bandwidth achieved by a slice from a given amount of resources varies with the channel quality of the users and the flow scheduling policies. H-CRANs have to consider resource sharing for uplink traffic, while UEs originate the uplink traffic randomly, which makes it hard to achieve the two conflicting goals of isolation and efficient resource utilization across slices. Meanwhile, H-CRANs often incur considerable overhead due to signaling and retransmissions that must be properly accounted for, which hurts resource utilization and flexibility.

Large-Scale Self-Organizing H-CRANs (LS-SON)

To fulfill the universal plug-and-play functionality, offload the traffic from the core network, and manage computing and radio resources more efficiently using the service-aware controls, the large-scale self-organizing functions are critical to guarantee the giant RRHs and ACEs work in an intelligent manner. SON was proposed to reduce operational costs for service providers in LTE cellular systems and HetNets [14]. Considering that too many parameters should be configured and optimized due to the combinations of heterogeneous convergence and cloud computing in H-CRANs, and that the radio resources are shared and virtualized, LS-SON is the key to integrate ultra learning, ultra planning, ultra configuration, and ultra optimization into a unified automatic process requiring minimal manual intervention with the centralization of cloud computing. The LS-SON reduces the complexity of managing co-channel interference in H-CRANs, and saves operational costs to all RRHs and ACEs. LS-SON is used to harmonize the entire network management approaches and improve the overall operational efficiency. On the other hand, the availability of LS-SON solutions leads to identifying powerful optimization strategies, and suppressing co-channel interference and improving EE and SE performance.

Due to the existence of centralized Node C co-located with SDN in the H-CRAN based 5G systems, the self-configuration, self-optimization, and self-healing functionalities are implemented in the centralized SON architecture. Since Node C

needs to converge multiple RANs and process the RRHs with cloud computing cooperatively, which is shown in Fig. 6, the inter-RAN and intra-RAN SON functionalities should be implemented in Node C.

For self-configuration, since RRHs are utilized to support transmissions with high data rates, the physical cell identifier (PCI) assignment is not necessary for RRHs, and only the radio resources should be self-assigned to each RRHs intelligently. However, the self-configuration of both PCI and radio resources should be fulfilled in ACEs. Thanks to Node C, these self-configuration cases for ACEs are handled in the centralized mode. For the self-optimization cases, energy saving and mobility load balance (MLB), mobility robustness optimization (MRO) are three key cases. To save the energy consumption, the dynamic energy saving strategy is used to optimize the number of hardware units in every moment, so that their number is minimized while the desired QoS is still ensured. In other words, RRHs and ACEs are switched off when they are not needed. The MLB has been envisaged as a distributed functionality, in which the algorithms are executed in the BS in 4G systems. However, the MLB will be located in Node C in a centralized mode. For self-healing, the perception that a problem is occurring or about to occur is the first step, followed by finding the root cause of the detected problem, and the final step is to identify and apply the appropriate corrective actions to (fully or partially, definitely or temporarily) restore the service. All these aforementioned LS-SON functions should be designed and fulfilled in the future.

Challenges and Open Issues in H-CRANs

Although there has been some progress and initial achievements in the above mentioned potential key techniques for H-CRANs, there are still many challenges ahead, including optimal resource allocation over the constrained fronthaul, energy harvesting, backward compatible standards development, and so on. Some classical challenges are discussed as samples in this section.

Fronthaul Constraints and Performance Optimization

Wired and wireless fronthauls are becoming increasingly important infrastructures in H-CRANs, which not only affects the data throughput available to UEs, but also determines the overall performance of the H-CRANs. Fronthaul with high bit rate and low latency enables larger-scale cooperative processing and networking, which in turn enables more efficient use of the available scarce resources. However, the practical fronthaul solutions have no sufficient end-to-end performances to meet the desired QoS requirements everywhere.

Considering that the non-ideal fronthaul may deteriorate overall SE and EE performance of H-CRANs, where both the limited capacity and time delay constraints are inevitable in practice, the solutions to tackle these constraints should be given in the future. To overcome the aforementioned problems related to the non-ideal fronthaul, new cooperative processing and resource allocation strategies need to be presented, and enhanced EE performance metrics that consider the constraints of fronthaul should be evolved. The optimal solution for resource allocation and admission control under the non-ideal fronthaul constraint is a non-deterministic polynomial-time (NP)-complete problem [15]. Even in moderate dense RRH deployment scenarios, computational/
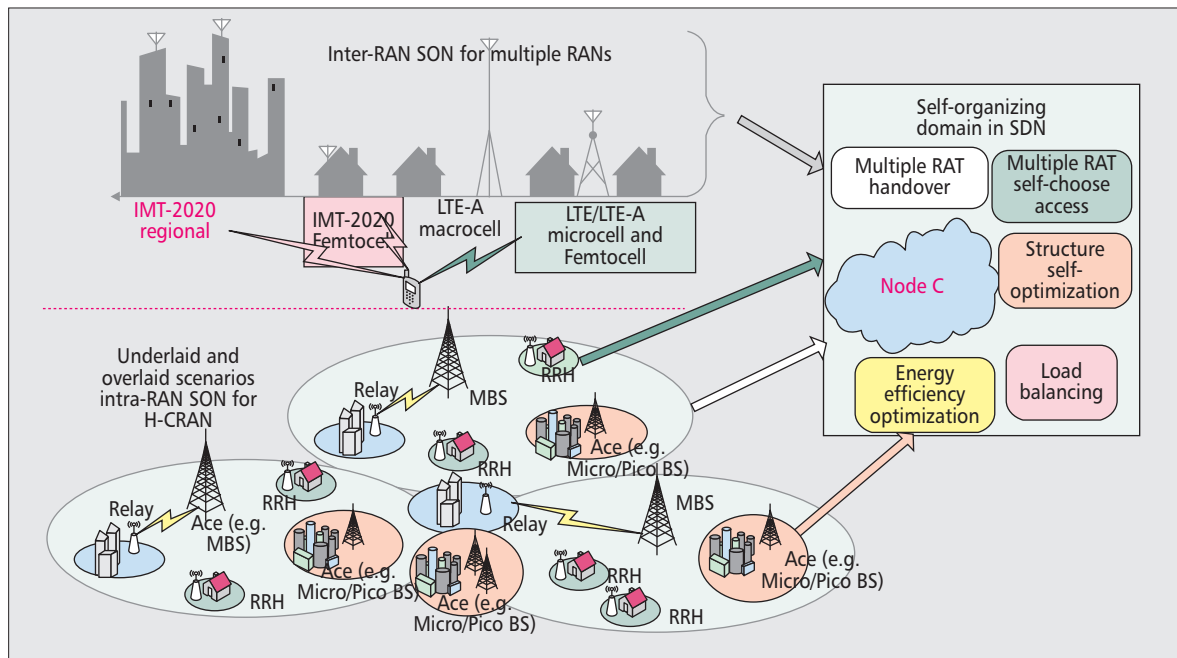
Figure 6. Self-organizing H-CRANs.

memory costs may prevent H-CRANs from achieving an optimal solution. Hence, the flexible iterative algorithms characterized by limited complexity should be designed to improve the overall SE and EE performances for H-CRANs in the future.

Energy Harvesting in H-CRANs

To save operational costs and decrease energy consumption, 5G systems are anticipated with renewable energy supply, such as solar and wind sources. Nevertheless, unlike conventional energy from the grid, renewable energy (e.g. harvested through solar panels or wind turbines) is intermittent in nature and can have different availabilities over time and space. Traditional methods such as energy storage with the use of capacity-limited and expensive batteries are far from enough for 5G systems to manage the fluctuations. To mitigate such uncertain renewable energy fluctuations and shortages, energy cooperation by sharing the excess energy among RRHs/ACEs with the help of Node C is a promising solution. However, one key problem that remains unaddressed to implement this cooperation among large-scale ultra dense RRHs/ACEs is how to motivate nodes with renewable energy to share excess energy with the others whose energy is insufficient, with some benefits in return. Furthermore, energy cooperation will occupy radio resources, which will sacrifice spectrum efficiency. To match peak-hour wireless traffic with limited spectrum, RRHs and ACEs can increase transmission power to improve spectrum efficiency, which indicates that energy harvesting should be adaptive to dynamic spectrum requirements and changeable traffic volumes.

Most existing research work for optimizing SE and EE performance assumes that transmission power is fixed and stable. However, based on energy harvesting, which has different availabilities over time and space, this work should be researched further. To summarize, SE and EE performance optimization in the ultra dense RRHs/ACEs with energy harvesting in H-CRANs is a promising and challenging area of research, where the transmission power of nodes should be adaptive to the packet traffic, radio channel fading, user's QoS, and offered energy.

Conclusion

In this article we have provided a summary of recent advances in the application scenario, system architecture, and key techniques for achieving high throughput and low energy consumption in heterogeneous cloud radio access networks (H-CRANs). To be compatible with the developments of SDN and NFV in the fifth generation wireless communication systems (5G), H-CRANs combine advantages of both HetNets and C-RANs to act as the access network. The key technologies, including advanced spatial signal processing, cooperative radio resource management, network function virtualization, and self-organizing network, are surveyed. Meanwhile, potential challenges and open issues are discussed in this article as well, including fronthaul constraints and their corresponding performance optimization, and energy harvesting. The presented key technologies and potential solutions to H-CRANs will be regarded as breakthroughs of the advanced next generation wireless communication systems.

References

[1] UMTS Forum Report 44, "Mobile Traffic Forecasts: 2010–2020 Report," UMTS Forum, Jan. 2011.
[2] J. G. Andrews et al., "What will 5G be?" IEEE JSAC, vol. 32, no. 6, Jun. 2014, pp. 1065–82.
[3] IMT.Vision, "Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," working document toward preliminary draft new recommendation ITU-R M.[IMT.Vision], Working Party 5D/TEMP/126 Rev. 1, Feb. 2013.
[4] M. Peng et al., "Ergodic Capacity Analysis of Remote Radio Head Associations in Cloud Radio Access Networks," IEEE Wireless Commun. Lett., vol. 3, no. 4, Aug. 2014, pp. 365–68.

[5] M. Peng et al., "Hierarchical Cooperative Relay Based Heterogeneous Networks," *IEEE Wireless Commun.*, vol. 18, no. 3, Jun. 2011, pp. 48–56.

[6] R. Irmer et al., "Coordinated Multipoint: Concepts, Performance, and Field Trial Results," *IEEE Commun. Mag.*, vol. 49, no. 2, Feb. 2011, pp. 102–11.

[7] M. Peng et al., "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.

[8] A. Lara et al., "Network Innovation Using Openflow: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, Feb. 2014, pp. 493–512.

[9] S. H. Park et al., "Joint Precoding and Multivariate Backhaul Compression for the Downlink of Cloud Radio Access Networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, Nov. 2013, pp. 5646–58.

[10] C. L. I et al., "Recent Progress on C-RAN Centralization and Cloudization," *IEEE Access*, vol. 2, Sep. 2014, pp. 1030–39.

[11] E. Larsson et al., "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 186–95.

[12] N. Saquib et al., "Fractional Frequency Reuse for Interference Management in LTE-Advanced HetNets," *IEEE Wireless Commun.*, vol. 20, no. 2, Apr. 2013, pp. 113–22.

[13] J. Li et al., "Resource Allocation Optimization for Delay-Sensitive Traffic in Fronthaul Constrained Cloud Radio Access Networks," *IEEE Syst. J.*, accepted by Oct. 2014.

[14] M. Peng et al., "Self-Configuration and Self-Optimization in LTE-Advanced Heterogeneous Networks," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 36–45.

[15] L. Zhou et al., "Uplink Multicell Processing with Limited Backhaul via Per-Base-Station Successive Interference Cancellation," *IEEE JSAC*, vol. 31, no. 10, Oct. 2013, pp. 1981–93.

Biographies

MUGEN PENG [M'05, SM'11] received the Ph.D. degree in communication and information systems from Beijing University of Posts & Telecommunications (BUPT), China, in 2005. He is a full professor with the School of Information and Communication Engineering in BUPT. His main research areas include cooperative communication, heterogeneous network, and cloud communication. He has authored/coauthored over 40 refereed IEEE journal papers and over 200 conference proceeding papers. He received the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in IEEE WCNC 2015, GameNets 2014, CIT 2014, ICCTA 2011, IC-BNMT 2010, and IET CCWMC 2009.

YONG LI [M'13] received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2009. He is currently an associate professor with the School of Information and Communication Engineering, BUPT. His current research interests include cooperative communications and next-generation wireless networks.

ZHONGYUAN ZHAO [M'15] is currently a lecturer with the Key Laboratory of Universal Wireless Communication (Ministry of Education) at Beijing University of Posts & Telecommunications (BUPT), China. He received his Ph.D. degree in communication and information systems and a B.S. degree in applied mathematics from BUPT in 2009 and 2014, respectively. His research interests include network coding, MIMO, relay transmissions, and large-scale cooperation in future communication networks.

CHONGGANG WANG [SM'09] received his Ph.D. degree from BUPT in 2002. He is a member of technical staff with InterDigital Communications, focusing on Internet of Things (IoT) R & D activities, including technology development and standardization. His current research interests include IoT, mobile communication and computing, and big data management and analytics. He is the founding editor-in-chief of *IEEE Internet of Things*, and he is on the editorial board of several journals, including *IEEE Access*.