Fog-Computing-Based Radio Access Networks: Issues and Challenges

Mugen Peng, Shi Yan, Kecheng Zhang, and Chonggang Wang

Abstract

An F-RAN is presented in this article as a promising paradigm for the fifth generation wireless communication system to provide high spectral and energy efficiency. The core idea is to take full advantage of local radio signal processing, cooperative radio resource management, and distributed storing capabilities in edge devices, which can decrease the heavy burden on fronthaul and avoid large-scale radio signal processing in the centralized baseband unit pool. This article comprehensively presents the system architecture and key techniques of F-RANs. In particular, key techniques and their corresponding solutions, including transmission mode selection and interference suppression, are discussed. Open issues in terms of edge caching, software-defined networking, and network function virtualization are also identified.

Compared to the fourth generation (4G) wireless communication system, the fifth generation (5G) wireless communication system should achieve system capacity growth by a factor of at least 1000, and energy efficiency (EE) growth by a factor of at least 10 [1]. To achieve these goals, the cloud radio access network (C-RAN) has been proposed as a combination of emerging technologies from both the wireless and information technology industries by incorporating cloud computing into radio access networks (RANs) [2]. C-RANs have come with their own challenges in the constrained fronthaul and centralized baseband unit (BBU) pool. A prerequisite requirement for centralized processing in the BBU pool is an interconnection fronthaul with high bandwidth and low latency. Unfortunately, the practical fronthaul is often capacity and time-delay constrained, which has a significant decrease on spectral efficiency (SE) and EE gains.

To overcome the disadvantages of C-RANs with the fronthaul constraints, heterogeneous C-RANs (H-CRANs) have been proposed in [3]. The user and control planes are decoupled in such networks, where high power nodes (HPNs) are mainly used to provide seamless coverage and execute the functions of the control plane, while remote radio heads (RRHs) are deployed to provide high-speed data rate for packet traffic transmission in the user plane. HPNs are connected to the BBU pool via the backhaul links for interference coordination. Unfortunately, H-CRANs are still challenging in practice. First, as location-based social applications become more and more popular, the traffic data over the fronthaul between RRHs and the centralized BBU pool surges with a lot of redundant information, which worsens the fronthaul constraints. Besides, H-CRANs do not take full advantage of processing and storage capabilities in edge devices, such as RRHs and "smart" user equipments (UEs), which is a promising approach to successfully alleviate the burden of the fronthaul and BBU pool. Moreover, operators need to deploy a huge number of fixed RRHs and HPNs in H-CRANs to meet requirements of peak capacity, which creates serious waste when the volume of delivery traffic is not sufficiently large. To solve such challenges, revolutionary approaches involving new RAN architectures and advanced technologies need to be explored.

Fog computing is a term for an alternative to cloud computing that puts a substantial amount of storage, communication, control, configuration, measurement, and management at the edge of a network, rather than establishing channels for the centralized cloud storage and utilization, which extends the traditional cloud computing paradigm to the network edge [4]. Note that it is also called edge cloud computing, which pushes applications, data, and computing content away from centralized points to the logical extremes of a network. In this article, we unify them as fog computing. Based on fog computing, the collaboration radio signal processing (CRSP) can be not only executed in a centralized BBU pool in H-CRANs, but also hosted at RRHs and even wearable "smart" UEs. To efficiently support and integrate new types of "smart" UEs, the on-device processing and cooperative radio resource management (CRRM) with a little distributed storing should be exploited. Meanwhile, from the viewpoint of mobile applications, UEs do not have to connect to the BBU pool to download the packet if the applications happen locally or the same content is stored in adjacent RRHs. Inspired by these characteristics of fog computing, to alleviate the existing challenges of H-CRANs and take full advantage of local caching, CRSP, and CRRM functions at edge devices, including RRHs and "smart" UEs, the fog-computing-based RAN (F-RAN) architecture is proposed in this article.

There are some apparent advantages in F-RANs, including the real-time CRSP and flexible CRRM at the edge devices, the rapid and affordable scaling that make F-RANs adaptive to the dynamic traffic and radio environment, and low burden on the fronthaul and BBU pool. Furthermore, the user-centric objectives can be achieved through the adaptive technique among device-to-device (D2D), wireless relay, distributed coordination, and large-scale centralized cooperation. To incorporate fog computing in edge devices, the traditional RRH is evolved to the fog-computing-based access point (F-AP) by being equipped with a certain caching, CRSP, and CRRM capabilities.

Mugen Peng, Shi Yan, and Kecheng Zhang are with Beijing University of Posts and Telecommunications.

Chonggang Wang is with the InterDigital Communications.

0890-8044/16/$25.00 © 2016 IEEE

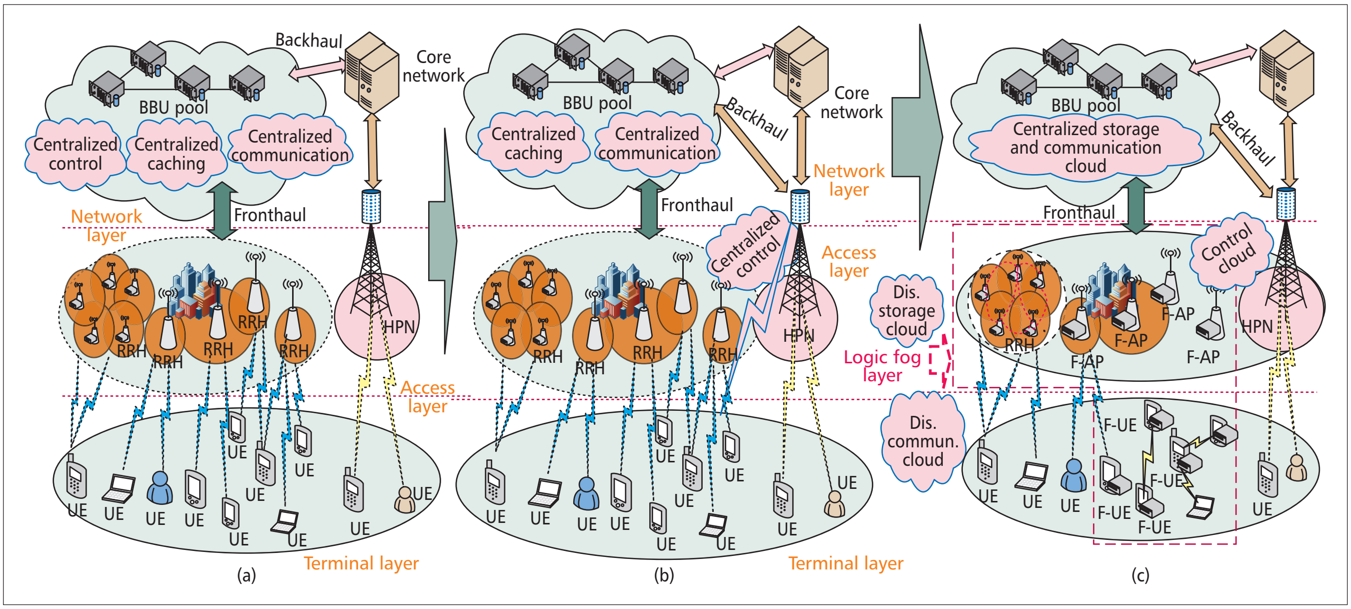

Figure 1. System architecture evolution through F-RANs: a) C-RAN architecture; b) H-CRAN artchitecture; c) F-RAN architecture.

Much attention has been paid to mobile fog computing and the edge cloud from the viewpoint of information sciences and the Internet of Things (IoT) recently. In [4], a fog computing platform to deliver a rich portfolio of new services and applications at the network edge has been proposed by Cisco first. The design of mobile fog as a programming model for large-scale latency-sensitive applications in the IoT has been introduced in [5]. Lately, in [6], an edge cloud and underlay network architecture has been proposed, and its corresponding simulation performance for edge caching has been presented. To the best of our knowledge, there is still no published work to discuss fog computing in 5G RANs.

In this article, we are motivated to make an effort to offer a comprehensive discussion on system architectures and technological principles in F-RANs. Specifically, the F-RAN system architecture is presented, where the new communication entity F-AP is defined and the software-defined F-RAN architecture is designed. Adaptive transmission mode selection and interference suppression in F-RANs are studied, and the corresponding performances are discussed. The future challenges and open issues are presented as well.

The remainder of this article is outlined as follows. F-RAN system architectures are introduced in the following section. Adaptive transmission mode selection is presented in the third section. Interference suppression techniques are discussed following that. Future challenges and open issues are then highlighted, followed by our conclusion.

F-RAN System Architecture

The F-RAN system architecture evolution from C-RAN is proposed in Fig. 1. In C-RAN and H-CRAN system architectures, all CRSP functions and application storage are centralized at the cloud server, which requires billions of UEs to transmit and exchange their data fast enough with the BBU pool. Note that the main difference between C-RANs and H-CRANs is that the centralized control function is shifted from the BBU pool in C-RANs to the HPN in H-CRANs. Meanwhile, some UEs can access the HPN to alleviate the burdens on the fronthaul of C-RANs. Furthermore, the two biggest problems in both C-RANs and H-CRANs are the long transmission delay and heavy burden on the fronthaul. A simple solution is to stop transmitting the entire torrent of data to the BBU pool, and process part of the radio signals at the local RRHs and even "smart" UEs. Meanwhile, to avoid all traffic being offloaded directly from the centralized cloud server, some local traffic should be delivered from the caching of adjacent RRHs (denoted by F-APs in F-RANs) or "smart" UEs (denoted by F-UEs) to save the spectral usage of constrained fronthaul and decrease the transmission delay.

As shown in Fig. 1c, some distributed communication and storage functions exist in the logic fog layer. Accurately, the proposed F-RAN takes full advantage of the convergence of cloud computing, heterogeneous networking, and fog computing. Four kinds of clouds are defined: global centralized communication and storage cloud, centralized control cloud, distributed logical communication cloud, and distributed logical storage cloud. The global centralized communication and storage cloud is the same as the centralized cloud in C-RANs, and the centralized control cloud is used to complete functions of the control plane and located in HPNs. The distributed logical communication cloud located in F-APs and F-UEs are responsible for the local CRSP and CRRM functions, while the distributed logical storage cloud represents the local storing and caching in edge devices.

The proposed system model for implementing F-RANs, illustrated in Fig. 2, can be regarded as a practical example implementing four kinds of clouds defined in Fig. 1c, comprising the terminal layer, network access layer, and cloud computing layer. Accurately, the F-APs and F-UEs in the terminal layer and network access layer formulate the fog computing layer. In the terminal layer, adjacent F-UEs can communicate with each other through the D2D mode or the F-UE-based relay mode. For example, F-UE13 and F-UE11 can communicate with each other with the help of F-UE12, in which F-UE12 can be regarded as a mobile relay. If there are some data to be directly transmitted between F-UE11 and F-UE12, the D2D mode is used. The network access layer consists of F-APs and HPNs. Similarly, with H-CRANs, all F-UEs access the HPN to obtain all system information related signaling, which fulfills the functions of the control plane. In addition, F-APs are used to forward and process the received data. F-APs are interfaced to the BBU pool in the cloud computing layer through the fronthaul links, while HPNs are interfaced to the BBU pool with the backhaul links, which indicates that the signals over fronthaul links should be large-scale processed in the BBU pool, and there is only coordination between the BBU pool and the HPN over the backhaul links. The traditional X2/S1 interface for the backhaul link is backward

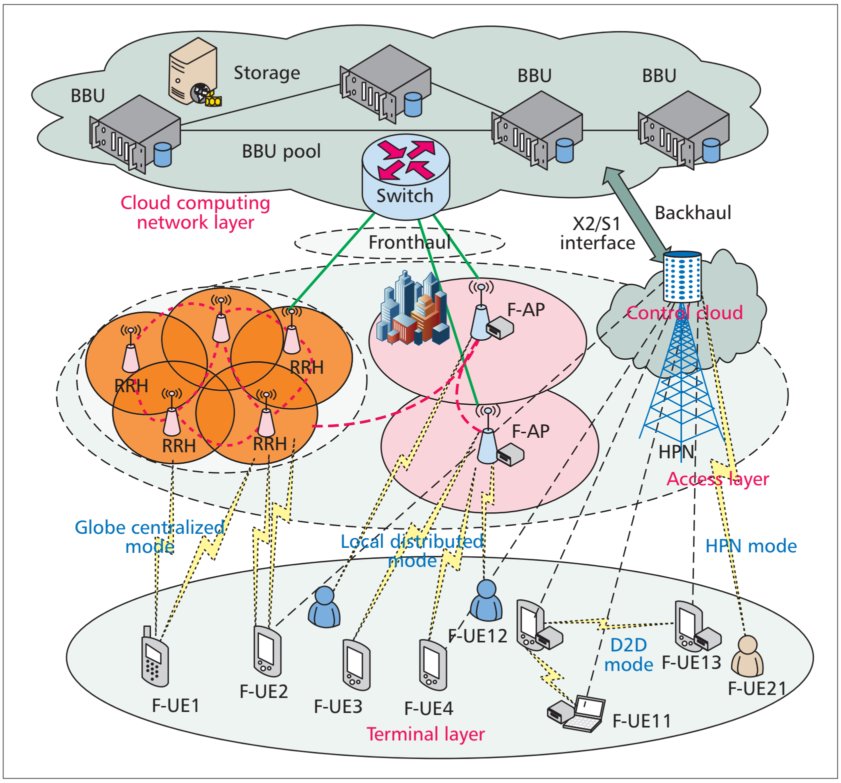

Figure 2. System model for implementing F-RANs.

compatible with that defined in Third Generation Partnership Project (3GPP) standards for Long Term Evolution (LTE) and LTE-Advanced systems. The BBU pool in the cloud computing layer is compatible with that in H-CRANs. Meanwhile, the centralized caching is located in the cloud computing layer. Since a large number of CRSP and CRRM functions are shifted to F-APs and F-UEs, the burden on the fronthaul and BBU pool are alleviated. Furthermore, the limited caching in F-APs and F-UEs can make some packet service offload from edge devices, and not from the centralized caching. The characteristics of F-RANs compared to C-RANs and H-CRANs are illustrated in Table 1.

Key Components: F-APs and F-UEs

To execute the necessary CRSP and CRRM functions in the logic fog computing layer adaptively, or offload the delivered packet traffic from edge devices locally, the evolved F-APs and F-UEs are presented to enhance the traditional RRHs and UEs, respectively. F-APs are mainly used to process local CRSP and CRRM for accessed F-UEs, offer interference suppression and spectral sharing for D2D F-UEs, and compress and forward the received information to the BBU pool through fronthaul. F-AP integrates not only the front radio frequency (RF), but also the local distributed CRSP and simple CRRM functions. By processing collaboratively among multiple adjacent F-APs, the overload of the fronthaul links can be released, and the queuing and transmitting latency can be alleviated. When all CRSP and CRRM functions are shifted to the BBU pool, F-AP is degenerated to a traditional RRH. F-UEs denote UEs accessing F-APs and working in the D2D mode.

Since the proposed F-RAN is evolved from HetNets and C-RANs, it is fully compatible with other 5G systems. The advanced 5G techniques, such as the massive multiple-input multiple-output (MIMO), cognitive radio, millimeter-wave communications, and non-orthogonal multiple access, can be used directly in F-RANs. The local distributed CRSP techniques among adjacent F-APs inherited from virtual multiple-input multiple-output (MIMO) can achieve high diversity and multiplexing gains for F-UEs without consuming fronthaul links. Herein, the interference among adjacent F-APs can be suppressed in a distributed manner under the help of coordinated multipoint (CoMP) transmission and reception. If the interference is not tackled efficiently by the local distributed CRSP and CRRM techniques, the global centralized CRSP and CRRM are triggered with the assistance of the BBU pool, which is the same way in C-RANs. When the traffic load is low, some potential F-APs fall into the sleep mode. While the traffic load becomes tremendous in a small special zone, F-APs and HPNs are active to absorb the high capacity, and the D2D or F-UE-based relay modes can be further triggered to meet the huge capacity requirement.

Hierarchical Architectures

The hierarchical architecture of F-RANs consisting of fog computing layer and cloud computing layer can make F-UEs adaptively work in the optimal mode. In the terminal layer, an F-UE can directly communicate with the adjacent F-UE in the D2D mode without the assistance of F-APs, in which the HPN is used to deliver overall control signaling for the D2D paired F-UEs. By reusing the same radio resources with F-UEs connecting with F-APs, the D2D mode is particularly beneficial to satisfy the demand of high data rate transmission, and also capable of enhancing the overall throughput. However, the D2D mode is severely constrained by the communication distance and the capability of F-UEs, in which traditional UEs without supporting D2D mode cannot be supplied. If the communication distance of two potential paired F-UEs is beyond the D2D distance threshold, the F-UE-based relay mode will be triggered to provide the communication for these two F-UEs with the third F-UE close to them.

In the network access layer, there are two types of edge communication entities: HPN and F-AP. Inherited from H-CRANs, the HPN is mainly used to deliver the overall control signaling and provide seamless coverage with basic bit rate for high mobile F-UEs. HPNs with massive MIMO are still critical to guarantee the backward compatibility with the existing wireless systems. The overall control channel overhead and cell-specific reference signals for F-RANs are delivered by HPNs, and thus F-RANs can decrease the unnecessary handover and alleviate the synchronous constraints. If the CRSP and CRRM functions are ended in F-APs, they have the same functions of small cell base stations, in which distributed interference coordination like CoMP is adopted to suppress the intra-tier and inter-tier interference. In addition, the adjacent F-APs are interconnected and formed into different kinds of topology to implement the local distributed CRSP, as shown in Fig. 3, in which each F-AP is connected to the others with the data and control interfaces S1 and X2, respectively. The interference among adjacent F-APs can be suppressed by the distributed CRSP and CRRM without the assistance of a BBU pool. In Fig. 3, the adjacent F-APs are formed into a mesh topology group to implement the distributed CRSP and deliver the packet traffic stored herein. In Fig. 3b, the tree-like topology for the F-AP connecting is illustrated. Both mesh and tree-like topologies can decrease negative influences of capacity-con-

Items	C-RANs	H-CRANs	F-RANs
Burden on fronthaul and BBU pool	Heavy	Medium	Low
Latency	High	High	Low
Decouple of user and control planes	No	Yes	Yes
Caching and CRSP	Centralization	Centralization	Mixed centralization and distribution
CRRM	Centralization	Centralization, and distribution between the BBU pool and HPNs	Mixed centralization and distribution
Performance gains	Fronthual constraint	Fronthual and backhaul constraint	Backhaul constraint
Implementing complexity	High in the BBU pool, low in RRHs and UEs	High in the BBU pool, low in RRHs and UEs	Medium in the BBU pool, F-APs, and F-UEs
Traffic characteristics	Packet service	Packet service, real-time voice service	Packet service, real-time voice service

Table 1. Advantage comparisons of C-RANs, H-CRANs, and F-RANs .

strained fronthaul links. Compared to the mesh topology, the wireless cluster feasibility in the tree-like topology is about 50 percent lower with significantly reduced network deployment and maintenance cost [7]. Therefore, the tree-like topology is preferred in practical F-RANs.

To achieve large-scale centralized CRSP and CRRM gains, F-APs are simplified into traditional RRHs, which forward received signals from UEs to the BBU pool. It is noted that the fronthaul constraints challenging C-RANs and H-CRANs are significantly alleviated in F-RANs because many CRSP and CRRM functions are executed in F-APs and F-UEs. Furthermore, much delivered packet traffic is stored not in the cloud server but in the edge devices.

The cloud computing layer is software defined, and is characterized by attributes of centralized computing and caching, which is inherited from C-RANs. All signal processing units work together in a large physical BBU pool to share the overall F-RAN's signaling, traffic data, and channel state information. When the network load grows, the operator only needs to upgrade the BBU pool to accommodate the increased capacity. The BBU pool is much easier for implementing the joint processing and scheduling to coordinate the inter-tier interferences between F-APs/RRHs and HPNs via the centralized large-scale CoMP approach. Centralized caching is used to store all global packet traffic.

Transmission Mode Selection

UEs access the F-RAN adaptively, and there are four transmission modes to be selected according to F-UEs' movement speed, communication distance, location, quality of service (QoS) requirements, processing, and caching capabilities: D2D and relay mode, local distributed coordination mode, global C-RAN mode, and HPN mode. In the D2D and relay mode, two F-UEs communicate with each other via the D2D or the UE-based wireless relay techniques. The local distributed coordination mode means that F-UEs access to the adjacent F-AP, and the communication is ended herein. The global C-RAN mode means that all CRSP and CRRM functions are implemented centrally at the BBU pool, which is the same as is done in C-RANs. Similarly, in H-CRANs, F-UEs with high movement speed or in the coverage hole of F-APs have to access the HPN, which is denoted by the HPN mode. As illustrated in Fig. 4, an adaptive mode selection is presented in this article to take full advantage of these four modes.

Except that all F-UEs periodically listen to the delivered control signaling, the optimum transmission mode is selected by the desired F-UE under the supervision of HPNs. To determine the optimal transmission mode for each F-UE, the movement speed of F-UEs and the distance of different F-UE pairs are estimated according to the public broadcast pilot channels from HPNs first. If an F-UE is in a high-speed mobile state or provides real-time voice communication, the HPN mode is triggered with high priority. If two F-UEs communicating with each other have a slow relative movement speed and their distance is not bigger than the threshold $D1$, the D2D mode

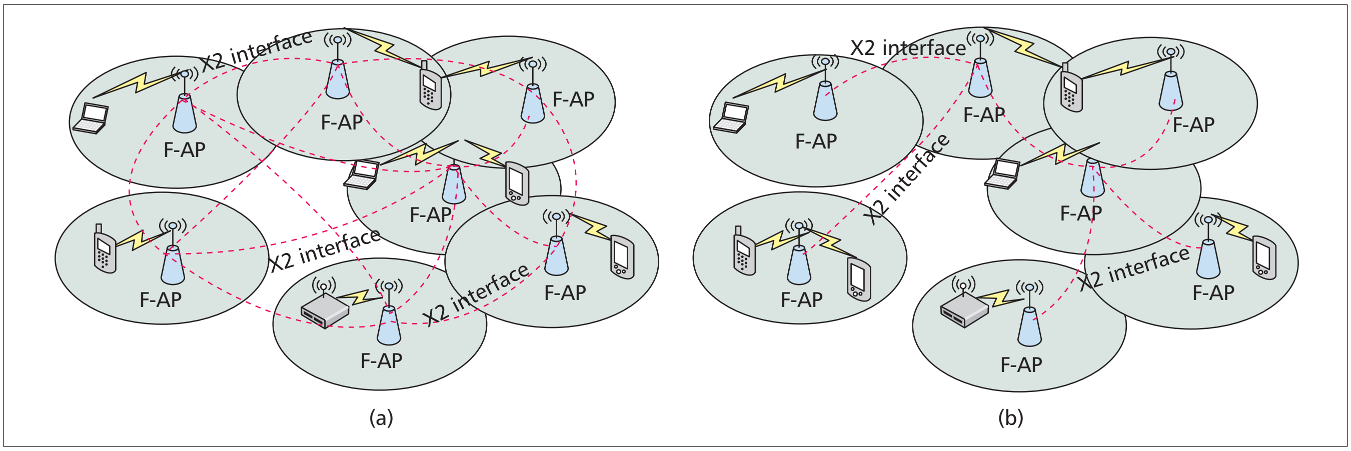

Figure 3. Two kinds of F-AP connecting topology in the network access layer: a) the mesh topology for the F-AP connecting; b) the tree-like topology for the F-AP connecting.

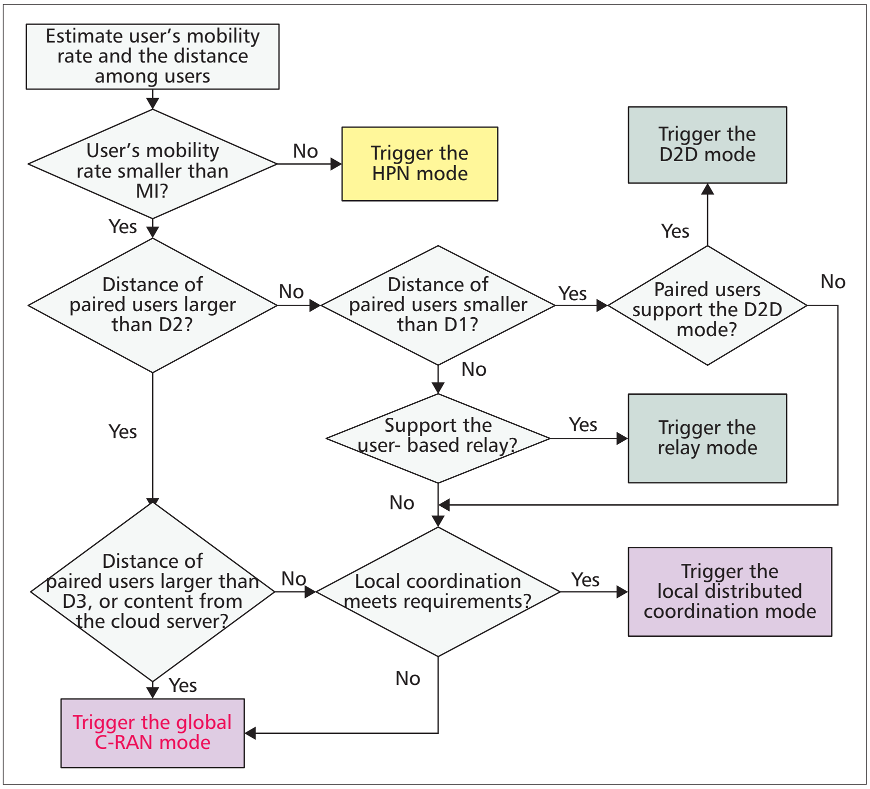
Estimate user's mobility rate and the distance among users
User's mobility rate smaller than MI?
No
Trigger the HPN mode
Yes
Distance of paired users larger than D2?
No
Distance of paired users smaller than D1?
Yes
Paired users support the D2D mode?
Yes
Trigger the D2D mode
No
No
Support the user- based relay?
Yes
Trigger the relay mode
Yes
No
Distance of paired users larger than D3, or content from the cloud server?
No
Local coordination meets requirements?
Yes
Trigger the local distributed coordination mode
Yes
No
Trigger the global C-RAN mode

Figure 4. Adaptive transmission mode selection in F-RANs.

is triggered. Otherwise, if their distance is bigger than $D1$ but smaller than $D2$, and there is one adjacent friendly F-UE acting as the F-UE-based relay for these two UEs to achieve better performance than the other modes, the F-UE-based relay mode is triggered. Furthermore, if two desired F-UEs move slowly, and their distance is larger than $D2$ but smaller than $D3$, or their distance is not bigger than $D2$, but at least one F-UE does not support the D2D and relay mode, the local distributed coordination mode is adopted. If the local distributed coordination mode cannot afford the expected performance, the distance between two desired F-UEs is bigger than D3, or the delivered content comes from the cloud server, the global C-RAN mode is triggered.

D2D and Relay Mode

The D2D and relay mode is a user-centric access strategy that makes communication only exist in the terminal layer, which can achieve significant performance gains benefitting from the D2D or wireless relay techniques and effectively relieve the burden on the fronthaul. In this mode, the HPN assigns the device identification for each F-UE. With necessarily lower antenna heights in D2D communication links, the fast fading channels are likely to contain strong line-of-sight components, which are different from the Rayleigh fading distribution in conventional wireless networks. In [8], the performance gains of D2D over cellular transmissions by taking into account the fading channel propagations with different Rician K-factors are analyzed and evaluated. As shown in Fig. 5, the numerical performance results of spatial average rate for both the cellular transmission and the D2D transmission are validated. It illustrates how the spatial average rate varies with the increase of the density of D2D users λ_D when the density of HPNs λ_M is fixed. In comparison to the Rayleigh fading channels, nearly 38 percent performance gains can be achieved when $K = 2$ dB in the low λ_D region, while nearly 68 percent performance gains are observed in the case of $K = 6$ dB. However, the spatial average rate performance is severely degraded for any scenario when the density of λ_D is sufficiently large. The performance gains from the F-UE-based relay mode can be referred to the two-way relay with network coding approaches in [9].

Local Distributed Coordination Mode

To decrease the burden on fronthaul and suppress the interference quickly, or directly offload the traffic not from the cloud server but from F-APs, the local distributed coordination mode is used, and the corresponding performance gains of interference coordination mainly source from CoMP. The F-AP cluster to execute the local distributed coordination mode is adaptively formed, which takes the implementing complexity and CoMP gains into account. The CoMP gains strictly depend on the topology of the F-RAN cluster and the backhaul capacity of the connecting F-APs. Based on the local distributed coordination mode, a remarkable increase of SE is achievable especially for the cell-edge user performance, which is on the order of 70 percent for the downlink and 122 percent in the uplink [6]. Meanwhile, the F-AP association is critical to improve SE. F-UE tries to access the optimal F-AP with the maximal received power strength when its minimum QoS can be satisfied. If the optimal F-AP does not have sufficient radio resource that can be allocated to the F-UE, or the interference to the adjacent F-APs is bigger than the predefined threshold, the F-UE tries to access the sub-optimal F-AP with allowances. It is noted that only a single F-AP is allowed to access each F-UE in this mode.

Global C-RAN Mode

If F-APs that should be coordinated to suppress the interference for the desired UE are not interconnected, or the content delivered to the desired UE is only stored in the cloud server, the global C-RAN mode is adopted. In this mode, RRHs

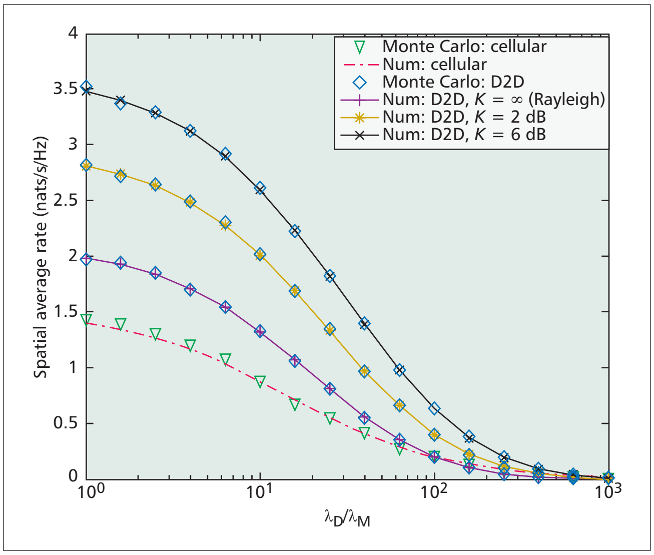

Figure 5. Spatial average rate for the cellular and D2D transmissions.

forward the received radio signals to the BBU pool, and the BBU pool executes all CRSP and CRRM functions globally and centrally, which is the same way as in C-RANs. Different from the local distributed coordination mode, several RRHs can work together for the desired UE to improve SE. With the help of the other three modes, the capacity demands on the fronthaul significantly decrease, which alleviates the capacity and latency constraints.

HPN Mode

If the content delivered to the desired UE is bursty with low volume, or the movement speed is beyond the predefined threshold, the HPN mode is preferred, which can decrease the overhead of the control channel and avoid unnecessary handover. This mode is mainly used to provide seamless coverage with basic QoS support. The resource sharing and performance gains of HPN mode are presented in [10], where an enhanced soft fractional frequency reuse (S-FFR) scheme can be used to mitigate the inter-tier interference between HPNs and F-APs. Only partial radio resources are allocated to UEs with low QoS requirements, and the remaining radio resources are allocated to F-UEs accessing F-APs with high QoS requirements. UEs with low QoS requirements accessing HPNs share the same radio resources with F-UEs accessing F-APs. The analysis and simulation results show that this S-FFR scheme has significant SE and EE gains in [10].

Interference Suppression

Since the same radio resources are shared among F-UEs with these four transmission modes, severe interference makes it challenging to improve the performance of F-RANs. Inspired by the CoMP in 3GPP, the interference suppression techniques in F-RANs can be categorized into coordinated precoding and coordinated scheduling. The coordinated precoding technique is utilized to decrease interference in centralized and distributed manners in the physical layer for the global C-RAN mode and the local distributed coordination mode, respectively. Coordinated scheduling is mainly used to suppress the interference in the medium access control (MAC) layer.

Coordinated Precoding

The coordinated precoding technique is generally categorized into the global and distributed manners. Global coordinated precoding includes massive MIMO for the HPN mode with large-scale antennas at a single site, and the large-scale cooperative MIMO for the global C-RAN mode with distributed F-APs located at different sites. The distributed coordinated precoding technique refers to joint processing CoMP with distributed F-APs in the same cluster for the local distributed coordination mode. To balance performance and complexity, the coordinated precoding size should be sparsely designed, in which only a small fraction of the overall entries in the channel matrix have reasonably large gains, and they can be ignored, leading to a great reduction in processing complexity and channel estimation overhead. In [11], the coordinated precoding cluster formation for F-APs is studied, and an explicit expression of the successful access probability for a fixed intra-cluster cooperation strategy is derived by applying stochastic geometry. By using the obtained theoretical result as a utility function, the problem of grouping F-APs is formulated as a coalitional formation game, and then the intra-cluster cooperation algorithm (called Algorithm 1 herein) based on the merge and split approaches is obtained. To estimate the performance gains, the grand cluster formation and no-clustering strategies, which can show the performance of the completely centralized and completely distributed schemes, respectively, are chosen as two baseline schemes. As shown in Fig. 6, the impact of the power consumption part is mitigated when $\tau = 0.1$, which can provide flexible choices for the cluster size settings. In this circumstance, the target data rate increases as the signal-to-interference-plus-noise ratio (SINR) threshold increases, and thus the average data rate keeps increasing in the lower and medium regions of γ_{th}. However, the successful access probability decreases as γ_{th} increases, and thus the increment of average data rate grows more slowly, or even declines in the high γ_{th} region. Since the power consumption is fixed, the trends of the EE curves almost match their corresponding average data rate curves.

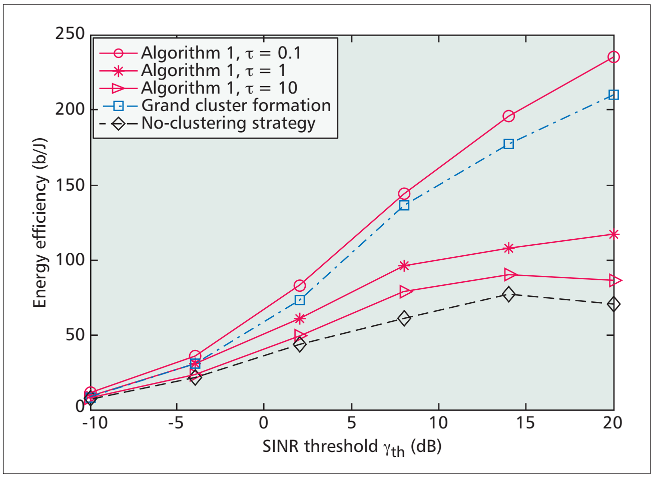

Figure 6. Energy efficiency of Algorithm 1. The energy exponent in the utility function is set as $\tau = 0.1$, 1, and 10, and the density of F-APs is $\lambda = 10^{-5}$.

Coordinated Scheduling

The coordinated scheduling is another emerging approach to mitigate interference in the MAC for the D2D and relay mode, local distributed coordination mode, and HPN mode. For example, to mitigate interference between D2D F-UEs and F-UEs accessing F-APs, centralized opportunistic access control (COAC) is presented in [8], where each D2D F-UE opportunistically accesses the sub-channels based on the centralized control of the HPN. The performance comparison between COAC and distributed random access control (DRAC) are shown in Fig. 7 for various spectrum occupation ratios of utilizing DRAC and COAC (i.e., ε). Since DRAC and COAC have the same effect on the D2D success probability, the performance is evaluated in terms of the cellular success probability, where sparse, medium, and dense D2D densities (denoted as $\lambda_D/\lambda_M = 10, 100, 1000$ with fixed λ_M) are considered. In addition, the asymptotic result is also plotted to show the tightness of the results. For comparison, the case of $\varepsilon = 0$ and $\varepsilon = 1$ are presented as the upper and lower bounds, respectively. It can be seen that the cellular success probabilities decrease with the increase of ε, and the extreme points exactly match with the upper and lower bounds that correspond to $\varepsilon = 0$ and $\varepsilon = 1$, respectively. Overall, COAC provides significant performance gains over DRAC by exploiting the benefits of centralized management and opportunistic access.

The optimal coordinated scheduling for F-RANs considering the cross-layer optimization of multiple objectives for the delay-aware circumstance is often complex because there is diverse interference among F-UEs with different transmission modes. To achieve the optimal solution taking the queuing delay into account, the equivalent rate constraint, Lyapunov optimization, and Markov decision process (MDP) are three common tools. The equivalent rate constraint approach converts the average delay constraints into equivalent average rate

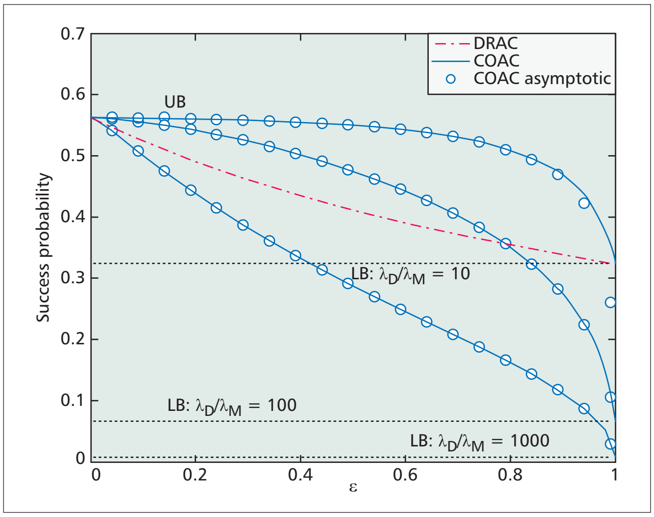

Figure 7. Cellular success probability with respect to ε under the DRAC and the COAC schemes. The upper bound is shown with the dashed line, while three lower bounds corresponding to λ_D/λ_M = 10, 100, and 1000 are provided with the dotted lines.

constraints using queuing theory or large deviation theory[12]. The Lyapunov optimization approach converts the average delay constraints into minimizing the Lyapunov drift-plus-utility function [13]. The Markov decision process approach is a systematic approach to solving the derived Bellman equation in a stochastic learning or differential equation manner. In [14], to minimize the queuing delay under the average power and fronthaul consumption constraints in the global C-RAN mode, the queue-aware rate and power allocation problem is formulated as an infinite horizon average cost constrained partially observed Markov process decision, which takes both the urgent queue state information and the imperfect channel state information at transmitter (CSIT) into account. A stochastic gradient algorithm is proposed to allocate power and transmission rate dynamically with low computing complexity and high robustness against the variations and uncertainties caused by unpredictable random traffic arrivals and imperfect CSIT.

Challenging Work and Open Issues

Although F-RAN is a promising technology to cope with the disadvantages in C-RANs and H-CRANs, there are still some challenges and open issues that remain to be discussed in the future, including edge caching, software-defined networking (SDN), and network functions virtualization (NFV) technologies.

Edge Caching

Caching aims to achieve a trade-off between the transmission rate and storage. If the transmission rate is high, the requirement for storage is low. In F-RANs, although the scale of content acquired by service providers is growing significantly, it is unnecessary to cache all content in the cloud server, which increases the end-to-end delay. Some local traffic is stored in edge devices, which can significantly decrease burden on the constrained fronthaul, and improve the performance of CRSP and CRRM. With caching and computation capabilities at edge devices, edge caching is applied to relax the traffic burden at the cloud server, and provide fast content access and retrieval at F-UEs [6]. Therefore, edge caching is a key component to improve the performance of F-RANs. The key benefits of edge caching in F-RANs include:

- The alleviation of burden on the fronthaul, backhaul, and even backbone
- The reduction of content delivery latency
- The flexibly implementation of object-oriented or content-aware techniques to improve F-RAN performance and user experience

Compared to the traditional centralized caching mechanism, the caching space at each F-AP and F-UE is practically small, and edge caching often has a low-to-moderate hit ratio, which is the probability that a requested content can be found in the cache. Hence, intelligent caching resource allocation strategies and cooperative caching policies among edge devices are mandatory. Meanwhile, to exploit the edge caching benefits, some key factors should be considered and jointly optimized, such as the cache load, cache hit ratio, number of content requests, cost of caching hardware, and cost of radio resource usage. Meanwhile, the caching policies, deciding what to cache and when to release caches in different edge devices, are crucial for improving the overall caching performance. The traditional caching policies, such as first-in first-out, least recently used, and least frequently used, should be evolved to appropriately improve the cache hit ratio in F-RANs.

Software-Defined Networking

Inherited from H-CRANs, F-RAN decouples the control and user planes, and the functions of F-APs can be reconfigured by the software stack with proprietary languages. Intuitively, since SDN decouples the control plane from the data plane via controllers and allows software to be designed independent of the hardware [15], F-RANs as the RAN of 5G systems can be seamlessly converged with SDN as the core network. With the CRSP and CRRM procedures incorporated into the edge devices in F-RANs, the SDN controlling, which is originally designed for wired core networks, can be extended to the physical layer in addition to the network layer, and more flexible and efficient network control can be achieved. For the highly flexible interfaces among different edge devices, SDN can be recognized as a basic enabler to achieve the flexibility and reconfigurability of F-RANs.

However, the data forwarding flow in SDN is mainly at the IP layer, and how to combine the functions of the MAC and physical layers for edge devices in F-RANs is still not straightforward. Meanwhile, SDN is based on the centralized manner, while the F-RAN has a high emphasis on the distributed manner for edge devices. Therefore, it is challenging to achieve SDN ideas in real F-RANs. SDN for F-RANs needs to define slices, which requires isolating the CRSP and CRRM in edge devices so as to provide non-interfering networks to different coordinators. The status and locations of edge devices should be reported to SDN promptly, based on which the SDN controllers can make decisions efficiently, which is also challenging because it will increase the burden on fronthaul and decrease the advantages of F-RANs. These aforementioned challenges are nontrivial and should be coped with for the successful rollout of F-RANs with SDN.

Network Function Virtualization

NFV is the concept of transferring the network functions from dedicated hardware appliances to software-based applications, which aims to revolutionize the telecommunication industry by decoupling network functions from the underlying proprietary hardware. Recently, academic researchers and network engineers are exploiting virtual environments to simplify and enhance NFV in order to find its way smoothly into the telecommunications industry [16].

Based on the SDN for F-RANs, the programmable connectivity between virtual network functions (VNFs) is provided and can be managed by the orchestrator of VNFs, which will mimic the role of the SDN controller. Furthermore, NFV can virtualize the SDN controller to run on the cloud server, which could be migrated to fit locations according to network needs. However, how to virtualize the SDN controller in F-RANs is still indistinct due to the distribution characteristic in edge devices. The security, computing performance, VNF interconnection, portability, and compatible operation and management with legacy RANs specified for F-RANs are major challenges and should be exploited in the future.

Conclusion

In this article, we have introduced a fog-computing-based radio access network architecture for 5G systems, which incorporates fog computing into H-CRANs. Compared to the traditional centralized cloud-computing-based C-RANs/H-CRANs, collaborative radio signal processing and cooperative radio resource management procedures in F-RANs are adaptively implemented at the edge devices, which are closer to the end users. With the goal of understanding further intricacies of key techniques, we have presented transmission mode selection and interference suppression. Within these two key techniques, we have summarized the diverse problems and corresponding solutions that have been proposed. Nevertheless, given the relative infancy of the field, there are still quite a number of outstanding problems that need further investigation. Notably, it is concluded that greater attention should be focused on transforming the F-RAN paradigm into edge caching, SDN, and NFV.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61361166005), the National Basic Research Program of China (973 Program) (Grant No. 2013CB336600), the State Major Science and Technology Special Projects (Grant No. 2016ZX03001020-006), and the National Program for Support of Top-Notch Young Professionals.

References

[1] M. Peng *et al.*, "Recent Advances in Underlay Heterogeneous Networks: Interference Control, Resource Allocation, and Self-Srganization," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2nd qtr., 2015, pp. 700–29.

[2] M. Peng *et al.*, "System Architecture and Key Technologies for 5G Heterogeneous Cloud Radio Access Networks," *IEEE Network*, vol. 29, no. 2, Mar. 2015, pp. 6–14.

[3] 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, Dec. 2014, pp. 126–35.

[4] F. Bonomi *et al.*, "Fog Computing and its Role in the Internet of Things," *Proc. Wksp. Mobile Cloud Computing*, Helsinki, Finland, Aug. 2012, pp. 13–16.

[5] K. Hong *et al.*, "Mobile fog: A Programming Model for Large-Scale Applications on the Internet of Things," *Proc. Wksp. Mobile Cloud Computing*, Hong Kong, China, Aug. 2013, pp. 15–20.

[6] Q. Li *et al.*, "Edge Cloud and Underlay Networks: Empowering 5G Cell-Less Wireless Architecutre," *Proc. Euro. Wireless 2014*, Berlin, Germany, May 2014, pp. 676–81.

[7] T. Biermann, *et al.*, "How Backhaul Networks Influence the Feasibility of Coordinated Multipoint in Cellular Networks," *IEEE Commun. Mag.*, vol. 51, no. 8, Aug. 2013, pp. 168–76.

[8] M. Peng *et al.*, "Device-to-Device Underlaid Cellular Networks under Rician Fading Channels," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, Aug. 2014, pp. 4247–59.

[9] M. Peng *et al.*, "Cooperative Network Coding in Relay-Based IMT-Advanced Systems," *IEEE Commun. Mag.*, vol. 50, no. 4, Apr. 2012, pp. 76–84.

[10] M. Peng *et al.*, "Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks", *IEEE Trans. Vehic. Tech.*, vol. 64, no. 11, Nov. 2015, pp. 5275–87.

[11] Z. Zhao *et al.*, "Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks", *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1207–21.

[12] C. Zarakovitis *et al.*, "Power-Efficient Cross-Layer Design for OFDMA Systems with Heterogeneous QoS, Imperfect CSI, and Outage Considerations," *IEEE Trans. Vehic. Tech.*, vol. 61, no. 2, Feb. 2012, pp. 781–98.

[13] R. Urgaonkar and M. Neely, "Opportunistic Cooperation in Cognitive Femtocell Networks," *IEEE JSAC*, vol. 30, no. 3, Apr. 2012, pp. 607–16.

[14] J. Li *et al.*, "Resource Allocation Optimization for Delay-Sensitive Traffic in Fronthaul Constrained Cloud Radio Access Networks," to appear, *IEEE Systems J.*

[15] S. Sezer *et al.*, "Are We Ready for SDN? Implementation Challenges for Software-Defined Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 36–43.

[16] H. Hawilo *et al.*, "NFV: State of the Art, Challenges, and Implementation In Next Generation Mobile Networks," *IEEE Network*, vol. 28, no. 6, Dec. 2014, pp. 18–26.

Biographies

MUGEN PENG (M'05, SM'11) received his Ph.D. degree in communication and information systems from Beijing University of Posts & Telecommunications (BUPT), China, in 2005. Now he is a full professor with the School of Information and Communication Engineering in BUPT. His main research areas include cooperative communication, heterogeneous networks, and cloud communication. He has authored/coauthored over 50 refereed IEEE journal papers and over 200 conference proceedings papers. He received the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and Best Paper Awards at IEEE WCNC 2015, GameNets 2014, IEEE CIT 2014, ICCTA 2011, IC-BNMT 2010, and IET CCWMC 2009.

YAN SHI is currently a Ph.D. candidate with the Key Laboratory of Universal Wireless Communication (Ministry of Education) at BUPT. His research interest focuses on the performance analysis and optimization of cloud- and fog-computing-based radio access networks.

KECHENG ZHANG is currently a Ph.D. candidate with the Key Laboratory of Universal Wireless Communication (Ministry of Education), BUPT. His research interest focuses on the radio resource allocation of fog-computing-based radio access 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 the *IEEE Internet of Things Journal* and on the Editorial Boards of several journals, including *IEEE Access*.