Fully Exploiting Cloud Computing to Achieve a Green and Flexible C-RAN

Jianhua Tang, Ruihan Wen, Tony Q. S. Quek, and Mugen Peng

The authors review the recent advances of exploiting cloud computing to form a green and flexible C-RAN from two cloud-based properties: centralized processing and the software-defined environment. For the centralized processing property, we include coordinated multipoint and limited fronthaul capacity, multicasting, and CSI issues in C-RAN. For the software-defined environment property, we summarize elastic service scaling, functionality splitting, and functionality extension.

ABSTRACT

By merging cloud computing into the RAN, C-RAN has been foreseen as a prospective 5G wireless systems architecture. Due to the innovative move of migrating the baseband processing functionalities to the centralized cloud baseband unit pool, C-RAN is anticipated to reduce energy consumption significantly to be a green RAN. Moreover, with the cloud-based architecture, lots of new functionalities and RAN designs are ready to be incorporated, which redefines the RAN as a flexible RAN. In this article, we review the recent advances of exploiting cloud computing to form a green and flexible C-RAN from two cloud-based properties: centralized processing and the software-defined environment. For the centralized processing property, we include coordinated multipoint and limited fronthaul capacity, multicasting, and CSI issues in C-RAN. For the software-defined environment property, we summarize elastic service scaling, functionality splitting, and functionality extension. We also include some of our recent research results and discuss several open challenges.

INTRODUCTION

Looking back on the mobile communications system's evolution, that is, from first generation (1G) (analog) through to 4G (LTE), the main efforts have been committed to obtaining faster data rates and lower latency. However, during this evolution, a side-effect is causing the system to look grim: the energy consumption. For instance, in 2012, more than 200 GW annual average power consumption in the information and communications technologies (ICT) industry was monitored, one quarter of which are from telecoms infrastructure and devices. To cut down on carbon emissions and have a sustainable future for the ICT industry, 90 percent reduction in energy consumption must be accomplished in the 5G era.

To reduce energy consumption, lots of research works have emerged from both the technique and infrastructure categories. To be more specific:

- In the technique category, people mainly focus on developing energy-efficient algorithms, including base station (BS) sleeping, cell zooming, multiplexing, beamforming, and so on.
- For the infrastructure category, most research attention is on producing energy-efficient hardware, including renewable-source-supported access points, energy-efficient radio heads, millimeter-wave backhaul, and so on.

Instead of upgrading or evolving from these two categories separately, the cloud radio access network (C-RAN) was proposed as a competitive 5G structure that combines the advances from both of them. There are three main components of a C-RAN: remote radio heads (RRHs), fronthaul links, and a baseband unit (BBU) pool. The key innovation of C-RAN is decoupling baseband processing functionalities from the RRHs and migrating these functionalities to the centralized cloud BBU pool, which consists of many general-purpose servers. Hence, the functionality at RRHs can be just as slim as basic signal transceiving.

What gives the C-RAN a big advantage over the conventional RAN in energy reduction is the centralized cloud BBU pool. Specifically, by cloud computing, the RAN manages the transitions not only from many distributed BSs to a centralized BBU pool, but also from a hardware-defined infrastructure to a software-defined environment. As a result, on one hand, cloud computing technology in the centralized BBU pool pushes the RAN to be more energy-efficient. For example, the cooling system is deployed for each BS in the conventional RAN; however, in C-RAN, the cooling system is deployed for the whole centralized BBU pool,[1] and the cooling power can be adaptive to the number of active servers, which is dynamically adjusted in cloud computing. On the other hand, the software-defined environment of the BBU pool facilitates more flexibility for the C-RAN. For instance, many new functionalities can be added onto a C-RAN by just upgrading the software, since the processing, controlling, and management in a C-RAN are all software-defined.

As a result, the energy consumption reduction objective in 5G turns out to be easier to fulfill. That is, much higher multiplexing gain can be obtained by the centralized cloud BBU pool, and renewable sources (e.g., solar power) are also appropriate for supporting RRHs [1]. In addition, compared to the the conventional RAN, the C-RAN offers more flexibility due to the software-defined cloud environment. All this means that a green and flexible C-RAN can be achieved

[1] Normally, there is no cooling system at the RRHs.

Digital Object Identifier:
10.1109/MCOM.2017.1600922

Jianhua Tang is with Chongqing University of Posts and Telecommunications and Seoul National University; Ruihan Wen is with the University of Electronic Science and Technology of China; Tony Q. S. Quek is with Singapore University of Technology and Design and Kyung Hee University; Mugen Peng is with Beijing University of Posts and Telecommunications. The corresponding author is Ruihan Wen.

0163-6804/17/$25.00 © 2017 IEEE

by exploiting cloud computing. Over the past several years, the research on cloud computing in C-RAN has been gone through two stages.

The First Stage: People make use of the centralized processing property brought by cloud computing, in which the cloud BBU pool is regarded as a central super node. That means this node has the global information of the whole system, and can manage and process most of the tasks. However, there is no detailed and in-depth study about the mechanisms inside this node.

The Second Stage: People begin to pay attention to another property, that is, the software-defined environment. At this stage, the detailed mechanisms of cloud computing in the BBU pool and how these mechanisms help improve the C-RAN's efficiency are investigated.

In this article, we discuss the recent advances on fully exploiting the cloud BBU pool to achieve a greener and more flexible RAN, from the above two stages (aspects): centralized processing and the software-defined environment. We also present some of our recent results and highlight some open challenges.

A comprehensive survey on C-RAN was recently conducted in [2]. Compared to [2], the contributions and significance of this article can be considered as follows:

- We emphasize how cloud computing works in C-RAN. This means that we look at C-RAN from the aspect of cloud computing instead of mobile communications. This viewpoint can provide more insights for C-RAN researchers.
- In addition, we unify lots of prospective applications that can be incorporated in C-RAN into the cloud-based service model: X as a service (XaaS).
- We include some recent academic results from our own research (e.g., elastic service scaling) and also some industry progress (e.g., big data as a service), neither of which is covered by [2].
- We try to systematize the RAN and core network in 5G by leveraging cloud computing. To achieve this, we discuss the decentralized core network as a service and C-RAN interacting with network slicing.

Centralized Processing

Centralized processing in the BBU pool offers a natural place to implement many theoretically mature techniques, and brings along new side-effects as well. In this section, we outline the problems that people have studied in C-RAN by leveraging the centralized processing property.

Coordinated Multipoint and Limited Fronthaul Capacity

Over the past decade, coordinated multipoint (CoMP) techniques have attracted comprehensive research attention from both academia and industry, which aim to enhance the system throughput, especially for the cell edge users. However, there are some challenges to be overcome to gain the high performance of CoMP, such as clustering and synchronization.

The emergence of C-RAN provides an ideal environment to implement CoMP, since most challenges in CoMP are eliminated by the centralized processing property of C-RAN (e.g., synchronization). In return, the system power consumption of C-RAN can be further reduced by leveraging CoMP. For instance, the set of active RRHs can be adjusted by the clustering algorithms in CoMP. Hence, the inactive RRHs can be switched into sleep mode to save power consumption. One typical CoMP technique is joint transmission, which duplicates users' desired data to multiple coordinated BSs and transmits the data to each user from multiple coordinated BSs simultaneously. Applying joint transmission in C-RAN means each user's data has to be shared among all the coordinated RRHs and their connected fronthauls. This causes a side-effect in C-RAN: the fronthaul's capacity becomes demanding. Therefore, considering the limited fronthaul capacity is imperative.

To reduce the data transmission rate in the fronthaul, another approach is performing compression in the fronthaul. For instance, the authors in [3] explore compression in C-RAN fronthaul uplinks. By leveraging the correlation between RRHs, they propose a joint decompression and demodulation algorithm in C-RAN fronthaul uplinks, that is, jointly conducting decompression and detection in a single step, to minimize the transmission rate on the fronthaul, and guarantee an acceptable distortion of the decompressed signal in the meantime.

Multicasting

The current RAN is designed to deliver information to specified individuals based on unicasting. However, in 5G, unicasting may not be a good choice anymore for some communication scenarios (e.g., live video streaming), from both the energy efficiency and spectrum efficiency aspects, especially when the fronthaul capacity is scarce.

Due to the centralized processing property, multicasting has been proposed as a promising replacement for unicasting in C-RAN for some communication scenarios. As a natural result, the transmit power consumption in C-RAN is foreseen to be reduced by utilizing multicasting, attributed to the multiplexing gain. For example, in [4], the authors investigate multicast beamforming for different user groups in C-RAN to minimize the weighted sum of backhaul cost and transmit power. Their results show a significant performance advantage by utilizing multicasting rather than unicasting.

CSI Issues

Due to the large number of RRHs and users' equipments (UEs), the channel state information (CSI) is becoming huge in C-RAN. Thus, the so-called curse of dimensionality effect will be a potential obstacle to enhance the performance of centralized processing. There are some works that look at reducing CSI overhead in C-RAN. For example, the authors in [5] propose a compressive CSI acquisition method, which only acquires the instantaneous channel coefficients for a subset of channel links and uses statistical CSI for the others. From the power consumption perspective, reducing CSI overhead means the power consumption to exchange CSI between RRHs and UEs can be cut down as well.

Owing to the centralized processing property, multicasting has been proposed as a promising replacement for unicasting in C-RAN for some communication scenarios. As a natural result, the transmit power consumption in C-RAN is foreseen a reduction by utilizing multicasting, attributing to the multiplexing gain.

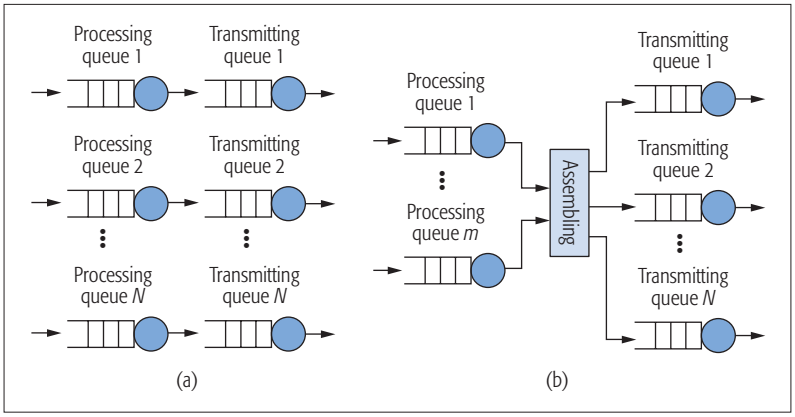

Figure 1. Two different elastic service scaling models: a) an idealized model.; b) a practical model.

Tackling imperfect CSI is always a big concern in wireless communications, and it becomes more significant for the centralized processing tasks in C-RAN, since even a slight imperfection in CSI between the estimated and true channel coefficients may lead to system-wide suboptimal operation. Recently, the authors in [6] make use of the stochastic optimization framework to deal with the noisy and delayed CSI in C-RAN. The results show that the impact of imperfect CSI can be reduced by their proposed approach.

Software-Defined Environment

Besides the centralized processing property, an innovative transition of C-RAN is that by introducing cloud computing, the BBU pool transfers from a hardware-defined infrastructure to a software-defined environment. In this section, we summarize some recent advances in C-RAN by utilizing the software-defined environment.

Elastic Service Scaling

Cloud computing has the famous five essential characteristics, that is, on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service, which have promoted it as a popular and successful computing paradigm over the past decade. By introducing cloud computing in the BBU pool, C-RAN also inherits these characteristics. For example, the rapid elasticity characteristic in C-RAN can be interpreted as meaning that, in the BBU pool, the operator can dynamically scale up and down the required computation resources to support baseband processing to improve the resource utilization. We also call this procedure in the BBU pool elastic service scaling.

Recently, we studied two different elastic service scaling models: an idealized model in [7] and a practical model in [8]. We extract the two main functionalities (i.e., baseband processing and signal transmitting) in C-RAN as two types of queues: the processing queue and transmitting queue, respectively.

The idealized model has the following highlights:

- One-to-one mapping, that is, one UE's incoming traffic is served by only one virtual machine (VM) in the BBU pool, and one VM only serves one UE's traffic. Hence, the number of VMs is equivalent to the number of UEs.
- To capture the elasticity, here, each VM's computation capacity can be dynamically adjusted according to incoming traffic rate, CSI, QoS requirement, and so on. This means that each VM's computation capacity is a variable to be determined.

The corresponding queueing model is represented in Fig. 1a, where the service rate of each processing queue is the computation capacity of each VM, and N is the number of UEs.

However, in a real system, it is not possible to implement the one-to-one mapping and adjust VMs' computation capacity for each UE individually in the BBU pool. Alternatively, in [8], we studied a more practical model, which has the following highlights:

- Multiple-to-multiple mapping, that is, one UE's incoming traffic can be served by multiple VMs in the BBU pool, and one VM can serve multiple UEs' traffic. Thus, the number of VMs is no longer equivalent to the number of UEs.
- Each VM's computation capacity is predefined and fixed, while the optimal number of active VMs is dynamically adjusted, that is, a variable to be optimized, according to the system status. This is how elasticity is captured here.

The practical model reflects the popular commercial cloud service models (e.g., Amazon EC2). The corresponding queueing system model is represented in Fig. 1b, where m is the optimal number of active VMs in the BBU pool.

For both models, there is a transmitting queue for each UE. The service rate of the transmitting queue is the wireless achievable rate to each UE, and the wireless achievable rate is also a variable to be identified. With these queueing system models, we are able to achieve a holistic design for C-RAN, with the following system delay constraint to couple BBU pool and RRHs:

delay in the processing queue + delay in the transmitting queue ≤ system delay threshold.

This constraint is applicable for both the idealized and practical models, but the delay terms have different mathematical expressions in different models.

An interesting trade-off is captured by the system delay constraint. Intuitively and qualitatively, more computation resource allocated in the BBU pool leads to lower delay in the processing queue, while resulting in higher power consumption in the BBU pool. However, thanks to the lower delay in the processing queue, a higher delay in the transmitting queue is hence acceptable based on the system delay constraint, which yields lower transmitting power consumption at the RRHs in return. This trade-off can be optimized mathematically to promote a greener C-RAN. More details and results are given in [7, 8].

Functionality Splitting

The most common architecture of C-RAN is a fully centralized system: processing, control, and management functionalities are located in the BBU pool, and the RRHs just keep basic RF functionality for signal transmission and reception. However, due to some practical constraints (e.g.,

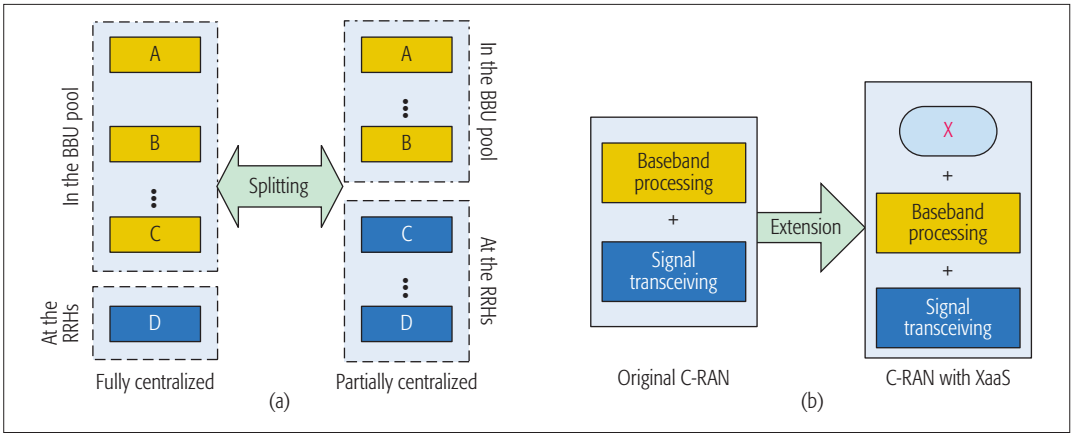

Figure 2. Cloud-assisted functionality flexibility in C-RAN: a) functionality splitting in C-RAN; b) functionality extension for C-RAN.

With the software-defined environment, the C-RAN operator can easily implement a functionality splitting system, instead of a fully centralized system. That means the operator is able to decide each function module to be realized in either the BBU pool or RRHs dynamically, based on different application scenarios.

limited fronthaul capacity), a fully centralized system may not be optimal in some scenarios.

With the software-defined environment, the C-RAN operator can easily implement a functionality splitting system instead of a fully centralized system. This means the operator is able to decide each function module to be realized in either the BBU pool or RRHs dynamically, based on different application scenarios (Fig. 2a).

As an example, Fig. 2a can be regarded as the functionality splitting of C-RAN's radio protocol stack [9]. The left side depicts a fully centralized scenario, in which modules {A, B, ..., C} denote {network management (NM), admission/congestion control (A/CC), radio resource management (RRM), media access control (MAC), and physical layer (PHY)}, and module D stands for the RF module. To reduce the fronthaul traffic amount, some modules can be migrated to the RRH side, as shown on the right side, where {A, ..., B} represents {NM, A/CC}, and {C, ..., D} represents {RRM, MAC, PHY, RF}. This results in a partially centralized C-RAN.

Functionality Extension

There are two main logical functionalities in original C-RAN: baseband processing and signal transceiving, where the baseband processing functionality is executed in the BBU pool, and the signal transceiving functionality is placed at the RRHs.

As the cloud-based BBU pool consists of many general-purpose servers, to fully utilize the software-defined environment of the BBU pool, it is reasonable and inevitable to incorporate more functionalities into it. As depicted in Fig. 2b, a new functionality, X, is added into C-RAN. We call this functionality extension C-RAN with XaaS. We list some recent progress and potential applications in XaaS in the following.

Mobile Cloud Computing as a Service (MCCaaS): Mobile cloud computing (MCC) is proposed to extend the computation ability and prolong the battery life of mobile devices by offloading the computation-intensive tasks to the cloud data center for processing. Although the cloud data center is much more powerful than a mobile device, conventional cloud data centers are always logically and physically far away from mobile devices, which lessens the benefits of MCC. Due to the proximity feature, the cloud-based BBU pool provides a new way to enjoy the advantages of MCC. That is, the computation-intensive tasks can be offloaded to the cloud-based BBU pool for processing instead of the conventional cloud data center in the remote end. In recent progress, the authors in [10] jointly study offloading, computation provisioning, and beamforming in C-RAN with MCCaaS.

Big Data as a Service (BDaaS): Different from MCCaaS, which aims to serve the individual user, BDaaS is proposed to serve the enterprise. For instance, Cazena, a startup company whose mission is to radically simplify enterprise big data processing in the cloud, is offering BDaaS. Under BDaaS, an enterprise's data infrastructure can be built, maintained, and upgraded nearly instantaneously, instead of spending months to set up an on-premises one. This is a very agile and cost-effective way of big data processing. By introducing BDaaS in C-RAN, many big-data-processing-based wireless communications applications can be executed in the BBU pool. An example is social-aware processing, where each user may belong to many different social groups, and each social group is classified by its interest, location, activities, and so on. Accurate and dynamic group classification is the basis to providing and improving specified wireless services for a targeted social group. This group classification problem is a typical application of big data processing, and resolving it in the BBU pool can greatly reduce the latency and backhaul traffic amount.

Decentralized Core Network as a Service: The core network plays the role of the brain of the whole network. It contains the data forwarding functionality in its user plane and the network controlling functionality in its control plane. Furthermore, the user plane and control plane are always coupled in the conventional core network. This core network is always logically located in the center of the whole network, which means all user traffic must go through it to access the Internet. However, in the 5G era, this centralized architecture will be overwhelmed by user data with much higher rate and lower latency requirements. Recently, SK Telecom proposed the concept of a "distributed core network," which separates the user plane

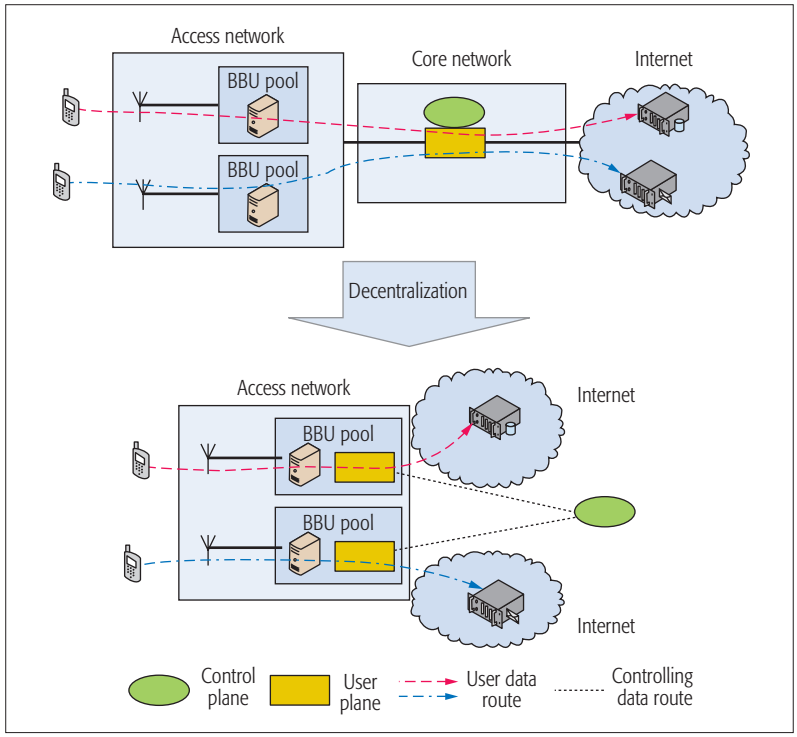

Figure 3. Evolution to the decentralized core network.

and control plane and partitions the physical core network into multiple virtual core networks, leveraging the software-defined network (SDN) and virtualization technology. In Fig. 3, we show the approach to alleviate the core network's burden by decentralizing the core network, with the assistance of C-RAN. Specifically, by migrating the user plane into the software-defined cloud BBU pool, users' data traffic through the physical core network can be greatly reduced.

Caching as a Service: Due to the limited space here, more details are elaborated in our previous work [11].

Thanks to the software defined environment, to realize XaaS in C-RAN is also not complicated. Most of the X can be accomplished by virtualization, without any additional hardware. For instance, MCCaaS can be achieved by generating some specified VMs (in the BBU pool, which support the applications running on the UE side. Then these VMs can help process the application tasks just as clones of the UE.

To sum up the last two sections, we qualitatively plot Fig. 4 to illustrate the interaction between greenness and flexibility. This means that although the two cloud-based properties (i.e., centralized processing and the software-defined environment) each has its own emphasis on greenness and flexibility, respectively, there is no hard bound between greenness and flexibility. In other words, some techniques can contribute to both greenness and flexibility in C-RAN (e.g., elastic service scaling).

Open Challenges

The study of fully exploiting cloud computing in C-RAN is the prerequisite to reach a greener and more flexible RAN. However, the state of the art is still not comprehensive and mature enough. We list some open challenges in this section.

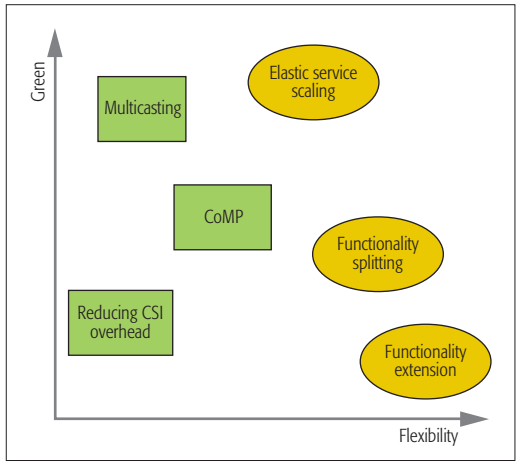

Figure 4. The interaction between green and flexibility.

The Two-Timescale Problem

In the last two sections, we summarize two main perspectives, centralized processing and the software-defined environment, that people have leveraged to exploit the cloud BBU pool. A straightforward extension is to utilize these two properties simultaneously. For example, in [12], we jointly study CoMP (due to centralized processing) and caching as a service (due to the software-defined environment).

However, these types of joint problems always have a common practical limitation: the two-timescale issue. Specifically, resource allocation in the BBU pool is always on the timescale of minutes to hours, while wireless resource allocation or beamforming is on the timescale of only milliseconds due to the wireless channel nature. As shown in Fig. 5, one slow timescale slot (e.g., 30 minutes)consists of T fast timescale slots (e.g., 1 ms). At the beginning of each slow timescale slot, we reallocate the resources in the BBU pool (e.g, activating new VMs), and at the beginning of each fast timescale slot, we redesign the beamformers based on the current CSI.

Problems under these two timescales are always coupled and complicated. For a typical example, the resources allocated in the BBU pool have to meet the wireless QoS requirements over a long time, that is, an entire slow timescale span in Fig. 5. However, this goal is not easy to achieve. This is because in order to achieve the goal at the time we perform resource allocation in the BBU pool, that is, the beginning each slow timescale slot, we have to consider every CSI in T fast timescale slots ahead, which is not possible (since we are unable to know the exact CSI in the future at that time). This becomes a main challenge to be resolved in our future work.

Reducing Latency

The main intention of making use of C-RAN is to improve the energy efficiency of conventional RANs. However, there is a well-known trade-off between energy consumption and latency in communication systems. Generally speaking, to reduce latency means to use higher-power transmitters and processors, which results in higher energy consumption. For example, this trade-off has been recently examined in edge cloud systems [13] and

device-to-device communications systems [14]. However, in C-RAN, there are extensive works aimed at improving energy efficiency, while only very few works pay attention to reducing latency.

In the 5G era, it is expected to have a much more stringent end-to-end latency requirement (i.e., 1 ms). As a potential 5G structure, C-RAN can contribute to reducing latency through the following potential approaches.

Functionality Splitting: In addition to reducing fronthaul traffic, functionality splitting is also applicable in reducing latency. To be more specific, Fig. 2a also illustrates the placement of atomic functions in the baseband processing structure. In particular, for some stringent latency requirement protocols, a fully centralized system may not be satisfiable. For instance, in LTE, the hybrid automatic retransmission request (HARQ) feedback only has 3 ms to be sent out once the corresponding frame is received. Therefore, relegating some atomic functions from the BBU pool to RRHs can effectively reduce the latency. Therefore, if the left hand side of Fig. 2a, {A, B, ..., C} denotes {coding, modulation, MIMO TX, IFFT}, and module D denotes the radio TX, after functionality splitting, the right side of Fig. 2a, {A, ..., B} can represent {coding, modulation, MIMO TX}, and {C, ..., D} can stand for {IFFT, radio TX} to reduce the latency (IFFT: inverse fast Fourier transform).

Fog RAN (F-RAN): To alleviate the traffic on C-RAN fronthaul and reduce latency, F-RAN is proposed as an evolution and complement of C-RAN. For example, for the Internet of Things (IoT) applications in 5G, some applications supported by C-RAN may not be able to meet their quality of service (QoS) requirements, such as low latency, high mobility, and location awareness, since the BBU pool is still relatively far from these "things." To overcome this limitation, one more tier of architecture needs to be deployed between UEs and RRHs. This tier is called fog, and consists of many edge devices. Hence, the RAN becomes a fog RAN. In F-RAN, edge devices, including RRHs and UEs, are also equipped with computation and storage capacities to perform radio signal processing, radio resource management, and caching. In this way, it is possible to transmit the entire torrent of data to the BBU pool via the fronthauls, and some of them can be just processed at the edge devices. As a result, the latency can be reduced accordingly. Nevertheless, F-RAN is not intended to replace C-RAN; It works cooperatively with C-RAN to extend the capability of C-RAN.

Caching as a Service: More details can be found in our previous work [11].

However, the investigation of leveraging C-RAN to achieve the 1 ms end-to-end latency requirement is still open.

Interacting with Network Slicing

Network slicing is proposed as a technique to construct virtual dedicated networks by logically separating the set of network functionalities and resources. These virtual dedicated networks are tailored by some special (technical or commercial) requirements. That means a virtual dedicated network can allow a group of UEs to exclusively access and use a part of the network functionalities and resources [15].

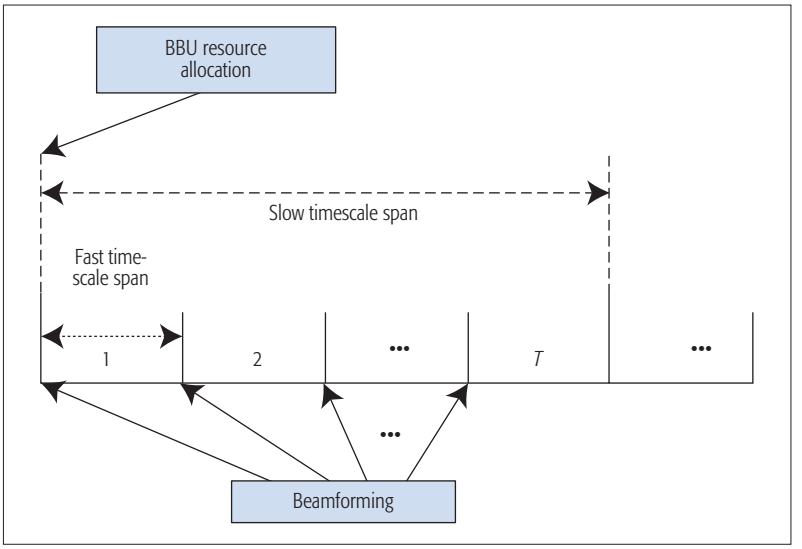

Figure 5. The two-timescale issue.

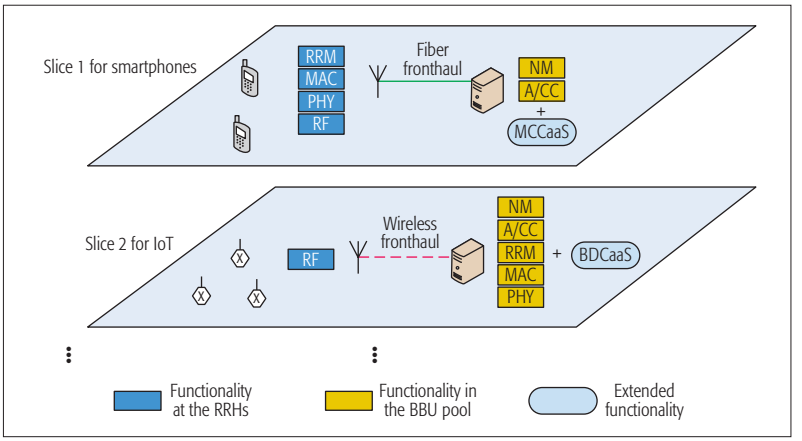

Figure 6. Network slicing for C-RAN.

Based on our discussion in the aforementioned sections, C-RAN is also ready to interact with network slicing. For example, in Fig. 6, we show two typical network slices in the coming 5G era. Slice 1 is for smartphones, which may contain some computation-intensive applications. Slice 2 is for IoT, where the sensors' data rate and latency requirements may not be high, but the features inside the data are very important. Based on these two different scenarios, the operator can produce two different slices in C-RAN accordingly.

Slice 1: To save the fronthaul overhead, we adopt the partially centralized structure for the radio protocol stack. As shown in the figure, only NM and AC/C are kept in the BBU pool; the rest are sunk to the RRHs. In addition, the fiber links are adopted as the fronthaul to further support the high data rate. Furthermore, we incorporate MCCaaS in the BBU pool to help execute the computation-intensive applications.

Slice 2: We still keep the fully centralized structure to coordinate the sensors better. Moreover, to save fronthaul cost and also keep the fronthaul data rate satisfactory, adopting the wireless fronthaul here can be a good choice. In addition, we include BDaaS in the BBU pool to analyze the features inside the data.

With the urgent need to reduce energy consumption in the ICT industry in the 5G era, C-RAN, a solution candidate, has been attracting a broad range of research attention. Due to its cloud-based architecture, C-RAN is going to be not only green but also flexible.

Therefore, together with network slicing in the core and aggregation network, we are able to accomplish the "network on demand" concept (proposed by AT&T). A fundamental of network slicing is that one slice should not be affected by the behavior of other slices. Nevertheless, this rule is difficult to accomplish in the sliced C-RAN due to "inter-slice" interference. As the investigation on network slicing is at a rudimentary level, to design, realize, and manage a C-RAN with multiple isolated slices, and to holistically slice the RAN, aggregation network, and core network are still big challenges.

Conclusion

With the urgent need to reduce energy consumption in the ICT industry in the 5G era, C-RAN, a solution candidate, has been attracting a broad range of research attention. Due to its cloud-based architecture, C-RAN is going to be not only green but also flexible. In this article, we dissect and review the benefits of cloud computing in C-RAN from two aspects, that is, centralized processing and the software-defined environment, which lay the foundation of a green and flexible C-RAN. We summarize some recently studied problems with respect to each aspect and introduce our latest results. Some potential research directions are also explored.

Acknowledgment

This work was supported in part by the Startup Funds of Chongqing University of Posts and Telecommunications under Grant A2016-114, the National Natural Science Foundation of China (NSFC) under Grant 61601071, the Open Foundation of State Key Lab of Integrated Services Networks of Xidian University under Grant ISN17-01, the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/01/2014, and the MOE ARF Tier 2 under Grant MOE2014-T2-2-002.

References

[1] A. Alameer and A. Sezgin, "Joint Beamforming and Network Topology Optimization of Green Cloud Radio Access Networks," *Proc. Int'l. Symp. Turbo Codes Iterative Info. Processing*, Brest, France, Sept. 2016, pp. 375–79.
[2] M. Peng *et al.*, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308.
[3] T. X. Vu, H. D. Nguyen, and T. Q. S. Quek, "Adaptive Compression and Joint Detection for Fronthaul Uplinks in Cloud Radio Access Networks," *IEEE Trans. Commun.*, vol. 63, no. 11, Nov. 2015, pp. 4565–75.
[4] M. Tao *et al.*, "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, Sept. 2016, pp. 6118–31.
[5] Y. Shi, J. Zhang, and K. B. Letaief, "CSI Overhead Reduction with Stochastic Beamforming for Cloud Radio Access Networks," *Proc. IEEE ICC*, Sydney, Australia, June 2014, pp. 5165–70.
[6] Y. Cai, F. R. Yu, and S. Bu, "Dynamic Operations of Cloud Radio Access Networks (C-RAN) for Mobile Cloud Computing Systems," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 3, Mar. 2016, pp. 1536–48.
[7] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-Layer Resource Allocation with Elastic Service Scaling in Cloud Radio Access Network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, Sept. 2015, pp. 5068–81.
[8] J. Tang *et al.*, "System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, May 2017, pp. 3371–84.
[9] P. Rost *et al.*, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 68–76.
[10] J. Cheng *et al.*, "Computation Offloading in Cloud-RAN Based Mobile Cloud Computing System," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[11] J. Tang and T. Q. S. Quek, "The Role of Cloud Computing in Content-Centric Mobile Networking," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 52–59.
[12] J. Tang, T. Q. S. Quek, and W. P. Tay, "Joint Resource Segmentation And Transmission Rate Adaptation in Cloud RAN with Caching as a Service," *Proc. IEEE SPAWC*, Edinburgh, U.K., July 2016, pp. 1–6.
[13] X. Guo *et al.*, "An Index Based Task Assignment Policy for Achieving Optimal Power-Delay Tradeoff in Edge Cloud Systems," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.
[14] M. Sheng *et al.*, "Energy Efficiency and Delay Tradeoff in Device-To-Device Communications Underlaying Cellular Networks," *IEEE JSAC*, vol. 34, no. 1, Jan. 2016, pp. 92–106.
[15] X. Zhou et al., "Network Slicing as a Service: Enabling Enterprises' Own Software-Defined Cellular Networks," *IEEE Commun. Mag.*, vol. 54, no. 7, July 2016, pp. 146–53.

Biographies

Jianhua Tang [S'11, M'15] received his B.E. degree in communication engineering from Northeastern University, China, in 2010, and his Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2015. He is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China. Currently, he is a research assistant professor at Seoul National University, Korea. His research interests include cloud computing, content-centric networks, and cloud RAN.

Ruihan Wen received her B.E. degree in communication engineering from Jilin University in 2010 and her M.E. degree in electronic information science and technology from the University of Electronic Science and Technology of China (UESTC) in 2013, respectively. She is now a Ph.D. student at the National Key Laboratory of Science and Technology on Communications in UESTC. Her research interests include resource management, network virtualization, and network slicing technologies in future networks.

Tony Q. S. Quek [S'98, M'08, SM'12] received his B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology. At MIT, he earned his Ph.D. in electrical engineering and computer science. He is a tenured associate professor with the Singapore University of Technology and Design. He is currently an Editor for *IEEE Transactions on Communications* and was an Executive Editorial Committee Member for *IEEE Transactions on Wireless Communications*.

Mugen Peng [M'05, SM'11] received his B.E. degree in electronics engineering from Nanjing University of Posts and Telecommunications, China, in 2000, and his Ph.D. degree in communication and information systems from Beijing University of Posts and Telecommunications (BUPT), China, in 2005. He is a full professor with the School of Information and Communication Engineering at BUPT. His main research areas include wireless communication theory, radio signal processing, and convex optimizations.