Elastic Resource Utilization Framework for High Capacity and Energy Efficiency in Cloud RAN

Dario Pompili, Abolfazl Hajisami, and Tuyen X. Tran

Current radio access network architectures, characterized by a static configuration and deployment of base stations, have exposed their limitations in handling the temporal and geographical fluctuations of capacity demand.

ABSTRACT

Current radio access network architectures, characterized by a static configuration and deployment of base stations, have exposed their limitations in handling the temporal and geographical fluctuations of capacity demand. Moreover, small cell networks have exacerbated the problem of electromagnetic interference and decreased the energy efficiency. Although there are some solutions to alleviate these problems, they still suffer from static provisioning of BSs and lack of inter-BS communication. Cloud RAN is a new centralized paradigm based on virtualization technology that has emerged as a promising architecture and efficiently addresses such problems. C-RAN provides high energy efficiency together with gigabit-per-second data rates across software defined wireless networks. In this article, novel reconfigurable solutions based on C-RAN are proposed in order to adapt dynamically and efficiently to the fluctuations in per-user capacity demand. Co-location models for provisioning and allocation of virtual base stations are introduced, and pros and cons of different VBS architectures are studied. Also, the potential advantages of VBS clustering and consolidation to support recently proposed cooperative techniques like cooperative multipoint processing are discussed.

INTRODUCTION

Over the last few years, the proliferation of personal mobile computing devices like tablets and smartphones, along with a plethora of data-intensive mobile applications, has resulted in a tremendous increase in demand for ubiquitous and high data rate wireless communications. The current practice to enhance data rate is to increase the number of base stations (BSs) and go for smaller cells to increase the band reuse factor. However, additional deployment and maintenance of a large number of BSs bring high inefficiencies due to excessive capital and operating expenditures. It has also been found that increasing the BS density or the number of transmit antennae will decrease energy efficiency (EE) due to the exacerbation of the electromagnetic interference problem and of the cooling requirements of cell site equipment [1].

On the other hand, the spatial distribution of users and the demand for capacity vary depending on the time of the day and week (the so-called *tidal effect*). In traditional cellular networks, each BS's spectral and processing resources are only used by the active users associated with that BS, causing idle BSs in some areas/times and oversubscribed BSs in others. The use of small cells is quite efficient in terms of power consumption as well as the utilization of spectral and processing resources when the capacity demand is high and evenly distributed in space. However, it becomes less so when the data traffic is low and/or uneven due to static resource provisioning and fixed power consumption. In this article, we discuss how the centralization of baseband units (BBUs), together with enabling virtualization of BSs while leveraging the paradigm of software defined wireless networking (SDWN), can be an effective way to address these challenges.

Cloud radio access network (C-RAN) is a new architecture for cellular networks where the BSs' computational resources are pooled in a central location; its main characteristics are:

- Centralized management of computing resources.
- Reconfigurability of spectrum resources.
- Collaborative communications.
- Real-time cloud computing on generic platforms.

C-RAN consists of three main parts:

1. Remote radio heads (RRHs) plus antennae, which are located at the remote site and are controlled by virtual BSs (VBSs) housed in centralized processing pools.
2. The BBU (VBS pool) composed of high-speed programmable processors and real-time virtualization technology to carry out the digital processing tasks.
3. Low-latency high-bandwidth optical fibers, which connect the RRHs to the VBS pool.

The communication functionalities of the VBSs are implemented (in software) on virtual machines (VMs) hosted over general-purpose computing servers that are housed in one or more racks of a small cloud data center. In a centralized VBS pool, as all the information from the BSs is resident in a common place, BSs can exchange control data at gigabit-per-second speed.

In this article, we propose a novel elastic resource utilization framework in which the

This work was supported by the National Science Foundation (NSF) under Grant No. CNS-1319945.

The authors are with the Department of Electrical and Computer Engineering, Rutgers University.

VBS size, RRH density, and transmit power can be dynamically changed to meet fluctuations in per-user capacity demand. This *elasticity* brings significant improvement in user quality of service (QoS) as well as efficiency in energy and computing resource utilization within the C-RAN paradigm. Our solution includes a *proactive* and a *reactive* component: the former anticipates the fluctuation in per-user capacity demand and provisions the VBSs in advance for a certain (limited) horizon; the latter monitors the VM utilization and triggers over- or underprovisioning alerts when there is a mismatch between the expected resource utilization and the actual observation. We explore innovative models for VBSs that capture the effect of computing resource contention (CPU, memory, network interface) among co-located VBSs in racks or servers in the data center. We discuss pros and cons of different architectures ranging from the traditional all-in-one VBSs (like legacy BSs) to split PHY- and medium access control VBSs (MAC-VBSs), which is more suited to exploit specific hardware characteristics, minimize computing resource contention, and maximize resource utilization. We also present the novel idea of a *VBS-Cluster*, in which we merge VBSs serving a cluster into a unit VBS-Cluster while the RRHs' antennae in each cluster act as a single coherent antenna array distributed over a cluster region, and discuss its advantages.

The rest of this article is organized as follows. We present the state of the art; we describe the idea of elastic VBS and explain the proposed resource provisioning and allocation models; we introduce the VBS-Cluster idea, and explore some advantages that can be achieved through the cooperation of VBSs within a cluster; and finally, we draw our conclusions.

State of the Art

Centralized management of computing resources (i.e., BS pooling) renders information global, and hence enables cooperative communication techniques at the MAC and PHY layers that were previously not implementable due to the strict throughput/latency inter-BS coordination requirements. Examples of MAC- and PHY-layer enhancements include joint flow scheduling and load balancing, collaborative spatial multiplexing, interference alignment and cancellation, and advanced mobility management. Although work has been done on the aforementioned cooperative communication techniques that can benefit from the C-RAN characteristics, research on enabling technologies for C-RAN itself is at a nascent stage, so there are only a few works in this area.

In [2], a partitioning and scheduling framework is proposed that is able to reduce the compute resources by 19 percent. In [3], the authors present a solution for small cells that reconfigures the fronthaul based on network feedback to maximize the amount of traffic demand. The authors of [4] propose the concept of cell zooming, where the cell size is adaptively adjusted according to traffic load, user requirements, and channel conditions. The authors of [5] introduce a reconfigurable backhaul scheme to allow for a flexible mapping between the BBUs and radio access units (RAUs); through real-world experiments, they show that their proposed solution improves RAN performance and decreases energy consumption. In [6], the authors propose a cross-layer resource allocation model in which they optimize the set of selected RRHs and the beamforming strategies at the active RRHs in order to minimize the overall system power consumption. In [7], the authors explore the trade-off between full centralization and decentralization of BBUs, and provide an overview of the challenges for fifth generation (5G) networks and why cloud technology will be a key enabler for such networks. In [8], the authors propose low-complexity three-stage group-sparse beamforming algorithms to minimize the network power consumption in C-RAN. The authors of [9] consider the coordinated transmission problem to minimize the downlink power in C-RAN; in order to serve each user, they determine a set of RRHs and the precoding vectors for the RRHs to minimize the total transmission power subject to the fronthaul capacity constraint.

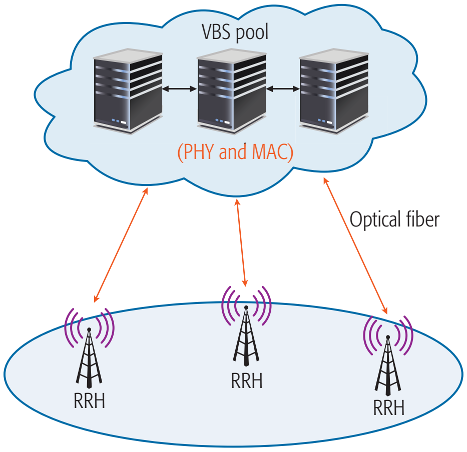

Figure 1. Cloud radio access network architecture, where the base stations are physically unbundled into virtual base stations and remote radio heads. Virtual base stations are housed in centralized processing pools and can communicate with each other at gigabit-per-second speeds.

In contrast to prior works on C-RAN, we propose the idea of elastic VBSs and dynamic RRH density that adapt to the fluctuations in capacity demand on the fly through demand-aware dynamic VM provisioning and allocation.

In contrast to prior works on C-RAN, we propose the idea of elastic VBSs and dynamic RRH density that adapt to the fluctuations in capacity demand on the fly through demand-aware dynamic VM provisioning and allocation. Moreover, we introduce the notion of VBS-Cluster, and propose innovative techniques where clustering, consolidating, and cooperation of VBSs improve the overall system performance.

Demand-Aware Provisioning

The number of active users at different locations varies depending on the time of day and week. This movement of mobile network load is referred to as the *tidal effect*. Today, each BS's spectral and computing resources are only used by the active users in that BS's cell. Deploying small cells for *peak traffic* (i.e., for the worst case) leads to grossly underutilized BSs in some areas/at some times and is highly energy inefficient; conversely, deploying for the *average traffic* leads to

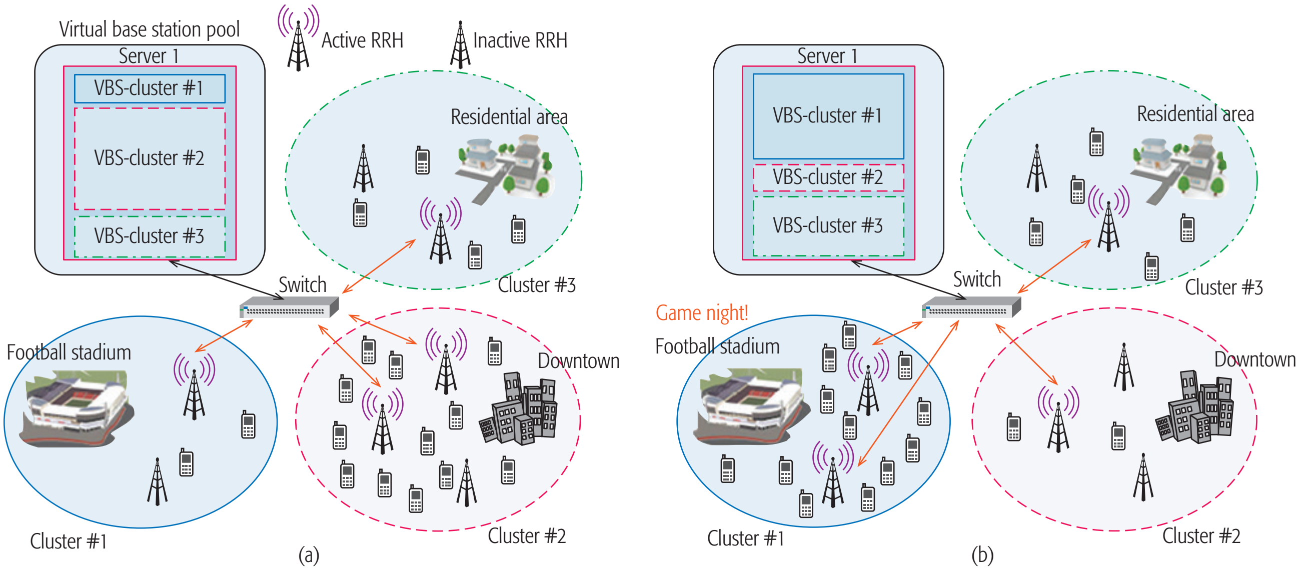

Figure 2. Virtualization in C-RAN allows for dynamic re-provisioning of spectral and computing resources (simplified here using different colored rectangles) to VBSs based on traffic demand fluctuation: a) and b) illustrate the movement of mobile network load from the downtown office area to the residential and recreational areas over the course of 24 hours, that is, during day and night, respectively; a) and b) also depict the corresponding changes in active RRH density and VBS size (note that active/inactive RRHs are identified by different icons, with or without wireless transmission).

oversubscribed BSs in some other areas/at other times. On the other hand, since traffic changes over time, there is no fixed cell size and transmission power that optimize the energy consumption; consequently, these system parameters can only be optimized for a fixed user density. Conversely, we propose a dynamic provisioning approach, at both the VBS and RRH sides, aimed at increasing the resource utilization and energy efficiency while providing a high level of QoS. As shown in Fig. 2, we cluster the neighboring RRHs and their corresponding VBSs, and change the density of active RRH and VBS size based on user density. Below we review the key features of our demand-aware provisioning approach.

Dynamic VBS resource provisioning: We advocate *demand-aware resource provisioning* in which VBSs are dynamically resized to meet the fluctuating traffic demand in the cellular network. As shown in Fig. 2a, during working hours, VBS-Cluster #2 will be provisioned with more computing resources compared to the ones serving a residential area (VBS-Cluster #3) or a stadium (VBS-Cluster #1). However, at night (Fig. 2b), the VBSs serving the stadium (e.g., on a game night) or a residential area will be provisioned with more resources than the ones downtown in order to meet the change in demand (VM resizing).

Size of a VM: All or some of the communication functionalities of a VBS (e.g., PHY, MAC, packet processing) are implemented in a VM. In order to achieve demand-aware dynamic VM provisioning, we introduce the notion of *size of a VM*, which is represented in terms of its processing power [CPU cycles per second], memory and storage capacity [bytes], and network interface speed [bits per second]. It primarily depends on the number of mobile users and type of data traffic (per-user capacity requirements) as well as on the computational complexity and memory footprint of the signal processing algorithms at the PHY layer, and the scheduling and frame processing algorithms at the MAC layer. In addition, the complexity of communication algorithms (e.g., the ones for inter-cell interference mitigation among densely deployed BSs in high-capacity RANs) may also affect the size of VMs. Therefore, to perform dynamic resource provisioning, a clear mapping from the number and type of mobile data users to the size of the VMs needs to be created. Hardware provisioning for VBSs must be such that the frame processing time be less than the frame deadline.

Known pattern vs. time-series prediction: Our solution for dynamic provisioning (or reprovisioning) of VBS resources to handle traffic fluctuations is composed of a *proactive* and a *reactive* component; in the former, the fluctuation in per-user capacity demand is predicted, and the computational resources are provisioned in advance for a limited time horizon. This anticipation is a result of *knowledge of known patterns* (e.g., day and night, weekdays and weekends, holidays, game schedules, etc.) or *predictions* based on advanced time-series analysis of historical traffic traces from the immediate as well as distant past. Once estimates of the number and combinations of different types of mobile data traffic are available, one just has to look up the closest profile and decide on the amount of resources to be provisioned for the VM.

Prediction uncertainties: Even though the proactive component allows for a smooth transition and greater optimization with respect to (w.r.t.) energy expenditure and resource utilization in the ensuing VM allocation procedure, it falls short in handling uncertainties. Some of the causes for uncertainties include unanticipated fluctuations in the number of users and per-user capacity demands in emergency scenarios arising out of natural (e.g., hurricanes, tsunamis) or man-made (e.g., industrial accidents, transportation system failures) disasters, unavailability of certain profiles, inaccuracies in the generated profiles, and mismatch between the generated profiles and reality due to hardware performance degradation. For these reasons, the reactive component monitors/profiles the CPU/memory/network utilization of the VMs and triggers over- or underprovisioning alerts when there is a "significant" mismatch between the expected resource utilization (based on the profile) and the actual observation.

A simple simulation scenario: To demonstrate the multiplexing processing gain (i.e., the increase in the region of feasibility) that can be achieved through dynamic resource provisioning, we simulated the following simple scenario: two BSs, one serving indoor users (or users in a downtown area with a large number of obstacles) and another serving outdoor users (say, a recreational area in a suburb). At each BS, we assume that each user's traffic belongs to one of the three following types with equal probability and in increasing order of priority:

- *Voice over IP* (very low and constant bit rate).
- *Light browsing* (bursty but low data rate).
- *Streaming/downloading* (high data rate).

We also assume that the cost of serving one user of each type at the downtown BS is higher than the corresponding cost at the suburb BS (this cost takes into account both the computational complexity as well as the memory footprint). We define *region of feasibility* as the total number of active users served by the BS pool with an acceptable blocking probability of 5 percent, which is a metric used in the context of voice calls. Here, for simplicity, we reuse the term to also convey an acceptable level of service degradation in data traffic. Figure 3 shows that dynamic resource provisioning (case 3) increases the region of feasibility (in terms of number of active users) by as much as 50 percent compared to the simplest static provisioning case (case 1). Note that knowledge of relative spatial distribution of users among BSs can help improve the feasibility region (case 2), but may also result in chronic over- and/or underprovisioning when the demand fluctuation is high. Greater benefits can be obtained when the distribution of users of different traffic types is unequal.

Dynamic RRH provisioning: Similar to what we mentioned above, deploying small cells (to provide enhanced spectrum resources for the peak traffic time) will make the network become energy inefficient due to the unavoidable energy costs when the capacity demand is low. For instance, circuitry, paging channel, cooling system, backhaul, and amplifiers all consume power so that even in a non-operational mode, BSs would consume a considerable amount of energy [10]. In traditional cellular networks, the cell planning and optimization, mobility handling, resource management, signal processing, and coverage are all done by each BS uniformly. In this case, even if the small cells have no traffic, they cannot be turned off [11]. Conversely, by decoupling BSs into VBS and RRH, the latter would only be responsible for providing spectral resources, and could be dynamically turned on and off as needed according to the traffic demand. Hence, in order to minimize energy consumption, we propose to optimize dynamically the number of active RRHs to adapt to the current traffic demand and user spatial distribution. For instance, as shown in Figs. 2a and 2b, due to the higher capacity demand during the day in cluster #2 (Fig. 2a), we provision it with more active RRHs than at nighttime (Fig. 2b) when we have lower capacity demand.

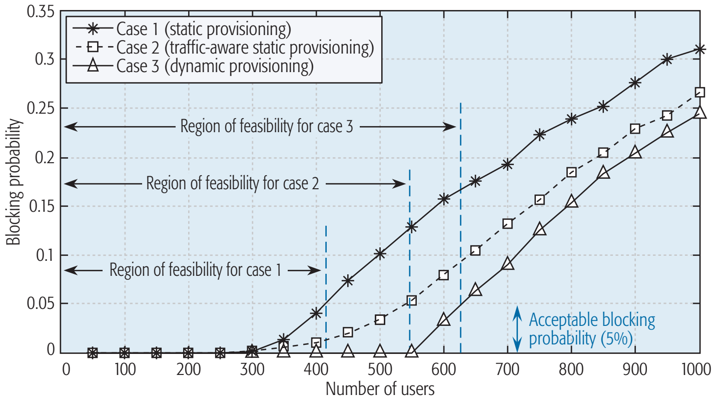

Figure 3. The benefit of dynamic computing-resource provisioning for VBSs at the remote data center: increase in the "region of feasibility" (w.r.t. active users).

Moreover, to minimize power consumption while ensuring a target data rate and user coverage, we need to adapt the transmission power of each cluster based on the density of its active RRHs. Both coverage and outage probability highly depend on the density and transmit power of the RRHs. This means that, given a fixed RRH density, we can minimize the transmit power of RRHs to target a certain coverage and outage probability. Since the RRH density of different clusters changes in time based on the capacity demand, we need to dynamically optimize the transmission power in each cluster. For instance, when the density of active RRHs becomes higher, each RRH has only a small coverage area, and users can receive acceptable signal-to-interference-plus-noise ratio (SINR) even when a lower output power is transmitted, which would save energy.

QoS-Aware VBS Allocation

Once the VMs holding the VBSs are provisioned, they have to be allocated to physical machines (PMs), that is, servers in the data center (called the centralized BS pool). The VM allocation has to be energy-, thermal-, and mobile-user-QoS-aware in order to fully realize the potential of C-RAN.

Thermal-aware VM consolidation: We advocate thermal-aware *VM consolidation* [12] for the VM-allocation problem. Thermal awareness, which is the knowledge of *heat generation* and *heat extraction* at different regions inside a data center, is essential to maximize energy and cooling efficiency as well as to minimize server system failure rates. Thermal-aware VM consolidation has the following three benefits:

1. The energy spent on computation can be saved by turning off the unused physical servers after VM consolidation.
2. The utilization of servers that are in the

In C-RAN, we are able to assign each cell to different clusters in order for them to cooperate with each other using different techniques. As associated VBSs of each cluster need a high-data-rate communication to perform cooperative techniques, they have to be allocated to the same server so to rely on high-speed inter-VBS connections.

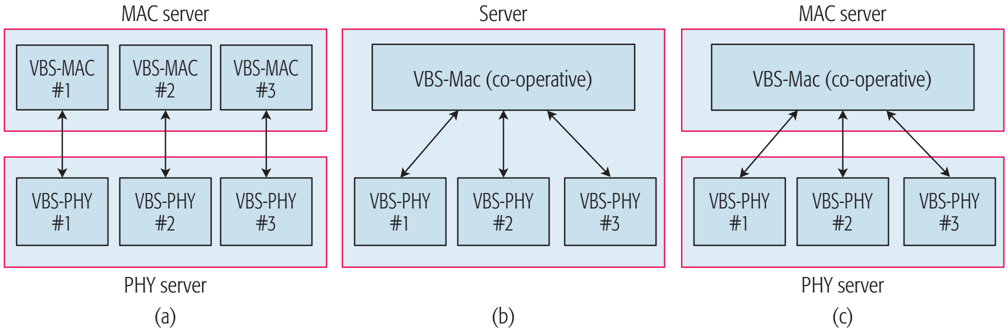

Figure 4. Split-VBS architectures: a) One-to-One mapping between PHY and MAC (different servers); b) Many-PHY-to-One-MAC (one server); c) Many-PHY-to-One-MAC (different servers). Note that the architectures in (a) and (c) (multi servers) can exploit the heterogeneity in datacenter server hardware.

"better cooled" areas of the data centers (i.e., with high heat extraction) can be maximized.

3. According to thermodynamics, heat can be extracted more efficiently (i.e., with a lower amount of work) by the cooling system from the consolidated server racks, which are hotter than non-consolidated racks.

In addition, consolidation on servers hosting VBSs allows efficient implementation of common functionalities such as signaling, channel state information (CSI) estimation for active users in a RAN, as well as for joint processing and scheduling techniques, such as ccordinated multipoint (CoMP) processing in 4G, for intercell interference mitigation.

Resource contention: Thermal and energy awareness alone, however, are insufficient for guaranteeing high VBS performance and maximizing energy and resource utilization efficiency. As multiple VMs share the same server resources (e.g., CPU, memory [RAM, cache], storage, and network interface), the performance of the corresponding VBSs in terms of per-user capacity and latency, and therefore the QoS of its mobile users, depend on the level of *contention for the computing resources* among co-located VMs. To factor in the effect of resource contention in VM allocation, we propose to classify the VBSs running a specific suite of algorithms for MAC- and PHY-layer functionalities as *CPU-, memory*-, and/or *network I/O-intensive*, and to develop co-location models that convey the degree of "compatibility" among co-located VMs. This way, we can incorporate the knowledge derived from co-location models into our VM allocation algorithm, thus making it QoS-aware.

Split-VBS architectures: To improve user QoS and resource utilization in C-RANs, we can deploy different architectures for VBSs. Figure 4 shows three possible/competing split-VBS architectures in addition to the traditional all-in-one VBSs in which the software modules for PHY and MAC are all implemented in one VM. The all-in-one architecture inherits characteristics from legacy BS designs, in which there is a one-to-one correspondence between MAC- and PHY-layer modules. *One of the primary motivations for this study is that the PHY and MAC layers are functionally quite different.*

One-to-one: PHY-layer processing requires vector execution techniques to accelerate signal processing, while MAC-layer processing requires multithread architecture and network accelerators for high-efficiency packet/protocol processing. In a data center with heterogeneous servers, exemplified as two separate servers in Fig. 4a (i.e., PHY server and MAC server), we can match the workload of BS-stack components with the capabilities of specific hardware.

Many-to-one: In general, communication between BSs can improve cellular system performance by exploiting the global and shared nature of information to make optimal decisions. For instance, in BS cooperation schemes, significant control information needs to be exchanged among neighboring BSs; however, cost, latency, and scarce interconnect capacity among BSs have been major impediments to the implementation of such schemes. We propose a split-VBS architecture, exemplified in Fig. 4b, in which the information of the MAC layers can be shared at gigabit-per-second speeds, making very-low-latency inter-BS communication possible. As a result, faster mobility management, more sophisticated interference suppression, and advanced cooperative multiple-input multiple-output (MIMO) techniques can be implemented to improve the user QoS. Finally, in order to take advantage of the heterogeneous processing pool as well as the high-speed inter-BS communication, we propose the architecture in Fig. 4c in which the VBS-MACs are merged together, and different physical servers are used for PHY- and MAC-layer processing.

Advantages of VBS Consolidation

In current distributed cellular systems, BSs can barely communicate with each other as the messages among BSs have to be exchanged through costly backhaul links. In C-RAN, as all the VBSs are located in a common rack of servers, they can exchange data with each other at gigabit-per-second speeds. Also, clustering the VBSs of the neighboring cells — together with enabling the coordination of the VBSs in the cluster — can greatly improve the system performance by exploiting the extra degrees of freedom, thus making optimal decisions.

We introduce the novel idea of a *VBS-Cluster*, according to which:

- All the VBSs associated with a certain cluster are merged together.

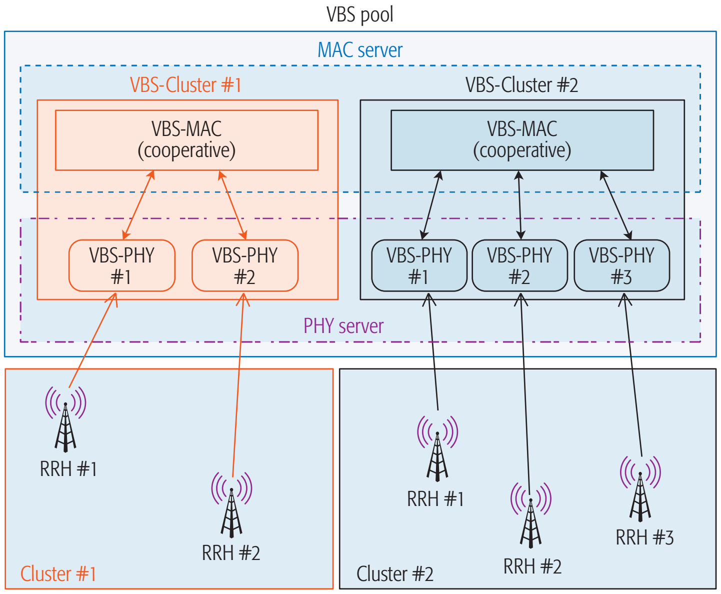

Figure 5. VBS-Cluster with a many-PHY-to-one-MAC architecture: VBSs associated with a cluster are merged together in the VBS-Cluster, and RRHs' antennae in each cluster act as a single coherent antenna array distributed over the cluster region.

- The RRHs' antennae in each cluster act as a single coherent antenna array distributed over the cluster region.

Figure 5 shows two VBS-Clusters, #1 (on the left) and #2 (on the right), where the sizes of the clusters are 2 and 3, respectively. Since in C-RAN VBSs are implemented on VMs, the size of VBS-Clusters (in terms of number of VBSs) can also be changed based on the network requirements. In such a case, the serving VBS-Cluster sends a CLS-REQ message to the target cluster to check whether it is ready to change the cluster size. As a response, the target cluster sends back a CLS-RSP message to the serving cluster to report whether it approves or rejects the CLS-REQ. If the decision is to change the size, the serving VBS sends a VBS-REQ message to the candidate cluster as a request to join. At this point, the target cluster acknowledges the VBS-REQ by sending a VBS-ACK to the serving VBS, which is finally added to the cluster.

In C-RAN, we are also able to assign each cell to different clusters in order for them to cooperate with each other using different techniques. As associated VBSs of each cluster need high-data-rate communication to perform cooperative techniques, they have to be allocated to the same server to rely on high-speed inter-VBS connections. Moreover, as the number of active users in the cluster determines the size of the VBS-Cluster, resource allocation needs to be performed for each cluster. We present here a few scenarios where clustering the VBSs to enable cooperation improves system performance.

Mobility management: In 4G wireless networks, only hard handover (HHO) (in which the connection between the serving BS and user is terminated before the connection between the new BS and the user is started) is defined to support users' mobility. As studied in [13], the service disruption time caused by HHO can be 250 ms or longer, which is intolerable for real-time services like voice over IP (VoIP). Note that with small cells, users perform handover more frequently, leading to a decrease in the perceived QoS; such degradation of QoS is a consequence of the short interruption in communication during HHO caused by overhead generated for controlling and managing the handover procedure itself. On the other hand, soft handover (SHO), which is a code-division multiple access (CDMA)-based handover scheme, can avoid service disruption as a user is actively connected to multiple BSs *simultaneously*. This contrasts with non-CDMA systems, in which a user can *only* be connected to one BS at a time. In C-RAN architectures, as the VBSs are co-located in a common place and can communicate and exchange data as well as controlling signals with each other, we are able to connect a user to multiple VBSs regardless of the modulation/access scheme. This means that we are able to use SHO for *both* non-CDMA and CDMA systems. By clustering VBSs, a user is actively connected to the associated RRHs as long as it remains in a certain cluster; in this case, a handover is needed less frequently (i.e., *only* when the user wants/needs to change the VBS-Cluster). To support CDMA in the VBS-Cluster, additional network resources are used; also, the associated VBSs need to perform a time correlation operation to detect the signal. On the other hand, to support a non-CDMA system, VBSs need to know the CSI from all the users, and matrix multiplications need to be performed to detect the signal.

Interference cancellation: In conventional cellular networks, each BS only serves users within its coverage area; thus, transmissions from neighboring cells interfere with each other, which decreases the SINR and spectral efficiency of cell edge users. A popular approach to address such interference is to employ CoMP, where neighboring cells are grouped together into clusters within which the BSs are connected to each other via the backhaul processing unit (BPU) [14]. In order to mitigate intra-cluster interference, the BSs in each cluster can perform coordinated beamforming and/or joint processing, which lead to improvements in spectral efficiency at the cost, however, of higher information exchange overhead among the BSs and more complex resource allocation. Although CoMP is able to reject the intra-cluster interference, it cannot mitigate the inter-cluster interference; consequently, cluster-edge users would still suffer from this type of interference. In addition, due to the distributed nature of the traditional cellular architecture, the latency and scarce interconnection capacity among the BSs have restrained the degree of cooperation among the BSs and the deployment of CoMP in practice.

These limitations can be overcome in C-RAN, where each cell can be associated with different clusters, and different clusters can communicate with each other at very high speeds. We envision a system employing CoMP over C-RAN to be highly capable of dynamically forming and reconfiguring user-centric clusters. In such a strategy, each scheduled user is in the center of its associated cluster, making it different from traditional static-clustering approaches where the cluster boundaries are fixed, and each cell belongs to one cluster only. This will eliminate cell-edge and

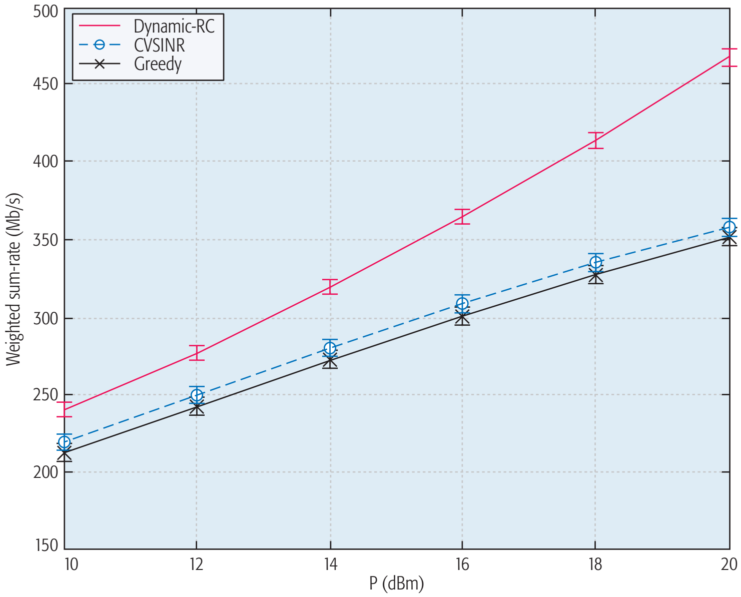

Figure 6. Improvement in downlink weighted sum rate (megabits per second) vs. RRH transmission power (dBm) of a C-RAN system employing dynamic clustering and cooperative beamforming. The competing strategies considered are dynamic-RC [15], a user-centric dynamic radio cooperation scheme; CVSINR, a user-centric heuristic clustering scheme with clustered virtual SINR beamforming; and greedy, a non-overlapping clustering scheme with zero-forcing beamforming.

cluster-edge users, mitigating both inter-cell and inter-cluster interference. The result of our work [15], shown in Fig. 6, demonstrates that a C-RAN system employing dynamic user-centric radio cooperation enables more effective beamforming techniques and outperforms traditional systems.

Technical challenges and open research issues: BSs have stringent real-time, low-latency, and high-performance requirements, to meet which the traditional virtualization technique is challenged. Specifically, in order to deploy a real-time VBS pool, the following requirements need to be met:

- Advanced real-time signal processing algorithms as well as high-performance low-power computing optimized for wireless signals.
- High-bandwidth, low-latency, low-cost BBU interconnection topology among physical processing resources in the baseband pool, which include the interconnection among the chips in a BBU, among the BBUs in a physical rack, and across multiple racks in the data center.
- Efficient and flexible real-time operating systems to achieve virtualization of hardware/resource management as well as dynamic allocation of physical processing resources to each VBS. This is needed to ensure latency and jitter control at the hardware level to support virtualization smoothly and efficiently.

Conclusion

We present novel reconfigurable solutions in the context of the cloud radio access network, a new centralized computing paradigm based on virtualization technology that has emerged as a promising architecture for broadband wireless cellular access. Such solutions adapt dynamically to fluctuations in per-user capacity demand, and offer higher energy efficiency and data rate (even in high-mobility scenarios). We advocate the need for co-location models for provisioning and allocation of VBSs, propose different VBS architectures, and discuss their pros and cons. Also, we present the advantages of VBS clustering, which can enhance energy efficiency and capacity in wireless cellular systems via advanced collaborative communication techniques.

References

[1] C. Li, J. Zhang, and K. Letaief, "Throughput and Energy Efficiency Analysis of Small Cell Networks with Multi-Antenna Base Stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, 2013, pp. 2505–17.

[2] S. Bhaumik *et al.*, "CloudIQ: A Framework for Processing Base Stations in a Data Center," *Proc. ACM MobiCom*, Aug. 2012, pp. 125–36.

[3] K. Sundaresan *et al.*, "FluidNet: A Flexible Cloud-Based Radio Access Network for Small Cells," *Proc. ACM MobiCom*, Sept. 2013, pp. 99–110.

[4] Z. Niu *et al.*, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Commun. Mag.*, vol. 48, no. 11, Nov. 2010, pp. 74–79.

[5] C. Liu *et al.*, "The Case for Re-Configurable Backhaul in Cloud-RAN based Small Cell Networks," *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1124–32.

[6] 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. 48, no. 9, 2015, pp. 5068–81.

[7] P. Rost *et al.*, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 68–76.

[8] Y. Shi, J. Zhang, and K. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, 2014, pp. 2809–23.

[9] V. N. Ha, L. B. Le, and N. Dao, "Energy-Efficient Coordinated Transmission for Cloud-Rans: Algorithm Design and Trade-Off," *Proc. IEEE CISS*, Mar. 2014, pp. 1–6.

[10] O. Arnold *et al.*, "Power Consumption Modeling of Different Base Station Types in Heterogeneous Cellular Networks," *Proc. FutureNetw*, June 2010, pp. 1–8.

[11] I. Chih-Lin *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.

[12] E. Lee, H. Viswanathan, and D. Pompili, "VMAP: Proactive Thermal-Aware Virtual Machine Allocation in HPC Cloud Datacenters," *Proc. IEEE HiPC*, Dec. 2012.

[13] W. Jiao, P. Jiang, and Y. Ma, "Fast Handover Scheme for Real-Time Applications in Mobile WiMAX," *Proc. IEEE ICC*, June 2007, pp. 6038–42.

[14] D. Lee *et al.*, "Coordinated Multipoint Transmission and Reception in LTE-Advanced: Deployment Scenarios and Operational Challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, Feb. 2012, pp. 148–55.

[15] T. X. Tran and D. Pompili, "Dynamic Radio Cooperation for Downlink Cloud-RANs with Computing Resource Sharing," *Proc. IEEE MASS*, Oct. 2015.

Biographies

Dario Pompili [SM] (pompili@cac.rutgers.edu) is an associate professor with the Department of Electrical and Computer Engineering at Rutgers University, where he directs the Cyber-Physical Systems Laboratory (CPS-Lab). He received his Ph.D. in ECE from the Georgia Institute of Technology in 2007. He previously received his Laurea (integrated B.S. and M.S.) and doctorate degrees in telecommunications and systems engineering from the University of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He is a recipient of the NSF CAREER '11, ONR Young Investigator Program '12, and DARPA Young Faculty '12 awards. He is a Senior Member of the ACM.

Abolfazl Hajisami (hajisamik@cac.rutgers.edu) started his Ph.D. program in electrical and computer engineering at Rutgers University in 2012. Currently, he is pursuing research in the fields of C-RAN, cellular networking, and mobility management under the guidance of Dr. Pompili. Previously, he received his M.S. and B.S. from Sharif University of Technology and Shahid Beheshti University, Tehran, Iran, in 2010 and 2008, respectively. His research interests are wireless communications, cloud radio access network, statistical signal processing, and image processing.

Tuyen X. Tran (tuyen.tran@cac.rutgers.edu) is working toward his Ph.D. degree in electrical and computer engineering at Rutgers University under the guidance of Dr. Pompili. He received his M.Sc. degree in electrical and computer engineering from the University of Akron, Ohio, in 2013, and his B.Eng. degree (Honors Program) in electronics and telecommunications from Hanoi University of Technology, Vietnam, in 2011. His research interests are in the application of optimization, statistics, and game theory to wireless communications and cloud computing.