Ad Hoc Networks 30 (2015) 128–143

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

Survey Paper

Dynamic provisioning and allocation in Cloud Radio Access Networks (C-RANs)

Dario Pompili *, Abolfazl Hajisami, Hariharasudhan Viswanathan

Department of Electrical and Computer Engineering, Rutgers University—New Brunswick, NJ, USA

ARTICLE INFO

Article history:
Received 21 October 2014
Received in revised form 23 January 2015
Accepted 21 February 2015
Available online 10 March 2015

Keywords:
Cloud Radio Access Network
Virtualization
Software Defined Radio
Cooperative multi-point processing

ABSTRACT

The Radio Access Network (RAN) is the most important part of a cellular wireless network. However, current cellular architectures have several disadvantages: they are not compatible with today's users' data-rate requests and do not leverage recent wireless enhancement techniques to achieve those data rates. Cloud Radio Access Network (C-RAN) is a new paradigm for broadband wireless access that provides a higher degree of cooperation and communication among Base Stations (BSs), in which all the BS computational resources are pooled in a central location, e.g., a set of physical servers in a datacenter. C-RAN represents a clean-slate design and allows for dynamic reconfiguration of computing and spectrum resources.

In this article, first explanations are provided on how this transformative paradigm can help overcome current cellular network limitations; then, its potential advantages to enable and support cooperative techniques like macro-diversity and collaborative spatial multiplexing are discussed. In addition, innovative C-RAN-based techniques to decrease the bandwidth-limiting Inter-Cell Interference (ICI) problem are proposed. Last, but not least, novel provisioning and allocation methods of Virtual Base Stations (VBSs) in the Base Band Unit (BBU) are proposed, and their pros and cons thoroughly discussed.

1. Introduction

The most important part of a cellular wireless network is the Radio Access Network (RAN) that provides wide-area wireless connectivity for Mobile Stations (MSs). In general, up to 80% of the Capital Expenditure (CAPEX) of a mobile operator is spent on the RAN [1]. In conventional RAN architectures, each Base Station (BS) only connects to a fixed number of sector antennae that cover a small area and only send/receive signals to/from the MSs in its coverage area. The hardware and processing equipment of each BS is located close to its antenna tower and there are no communication links connecting the BSs. Physical links only exist between BSs and their corresponding access network gateway. Hence, control messages between the BSs have to travel through costly backhaul links, and often even over a one-level higher layer in the aggregation hierarchy. The latency and scarce interconnect capacity among BSs have resulted in limited BS cooperation in practice. However, emerging wireless technologies such as *cooperative Multiple-Input Multiple-Output (MIMO)* or *coordinated scheduling and beamforming* require close cooperation among BSs.

Moreover, 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

* Corresponding author.
E-mail addresses: pompili@cac.rutgers.edu (D. Pompili), hajisamik@cac.rutgers.edu (A. Hajisami), hari_viswanathan@cac.rutgers.edu (H. Viswanathan).

http://dx.doi.org/10.1016/j.adhoc.2015.02.006
1570-8705/

in demand for ubiquitous and high data-rate wireless communications [2]. However, the system capacity is limited by the interference, which makes it difficult to improve the spectral efficiency and consequently data rate. To solve this problem, the fourth generation (4G) cellular communication system with peak downlink data rate of 1 Gbps has been envisioned. Long Term Evolution (LTE) systems based on Orthogonal Frequency Division Multiple Access (OFDMA) represent a major breakthrough in terms of achieving downlink peak data rates of 300 Mbps [3]. However, cooperative schemes used in LTE to increase the spectral efficiency cannot be fully deployed in current cellular networks for the reasons discussed above. Hence, LTE systems do not match yet the International Mobile Telecommunications Advanced (IMT-Advanced) "True 4G" requirements and a significant effort is being made towards the development of LTE-Advanced. For instance, Cooperative Multi-Point (CoMP) transmission and reception is one of the promising techniques being developed for LTE-Advance [4]. In CoMP, a set of neighboring cells are grouped into clusters, each consisting of connected BSs that share Channel State Information (CSI) and MS signals. This scheme allows for joint processing among BSs that can effectively mitigate the Inter-Cell Interference (ICI) and thus improve the spectral efficiency. However, the current cellular architecture has some drawbacks in terms of distributed and limited processing resources at the BSs as well as capacity constrained backhaul links that make it difficult to fully exploit the benefits of CoMP and consequently cluster-edge users still suffer from low Signal-to-Interference-plus-Noise Ratio (SINR).

Cloud Radio Access Network (C-RAN) [1,5] was introduced recently as a new paradigm for broadband wireless access that provides a higher degree of cooperation and communication among BSs. This architecture represents a clean-slate design and allows for dynamic reconfiguration of computing and spectrum resources. In C-RAN, all the BSs' computational resources are pooled in a central location, e.g., a set of physical servers in a datacenter, enabling communication among BSs with low latencies and exchange data at Gbps speeds. This centralization feature makes C-RAN flexible and suitable to support cooperative techniques such as joint scheduling, beamforming, and interference mitigation. For instance, as different clusters can communicate with each other, the inter-cluster interference can be mitigated via CoMP. In addition, in C-RAN, it is easier to dynamically adjust the cluster size and apply optimal resource allocation strategies so to improve the system capacity and energy efficiency. Furthermore, virtualization technology allows the implementation of Virtual Base Stations (VBSs) in a datacenter cloud, which will result in the reduction of Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) for operators.

In this article, firstly in Sections 2–4, we provide a comprehensive survey on C-RAN and describe its technical challenges. Then, we propose some C-RAN-based solutions to address the relevant open research issues. In Section 2, we investigate the shortcomings of current distributed cellular wireless systems. Then, in Section 3, we present C-RAN as a new centralized architecture and explain how it addresses the shortcomings of existing cellular systems. In Section 4, we study the implementation of VBSs and discuss the existing challenges and computational requirements of a VBS. In Section 5, we focus on resource provisioning and allocation strategies of VBSs in the centralized resource pool, and propose reactive and proactive provisioning schemes to improve the resource utilization efficiency and system performance. In Section 6, we explore the advantages of BS pooling and study the potential of C-RAN to improve cooperative techniques. We also present the idea of "*VBS-Cluster*", in which we merge VBSs serving a cluster into a unit VBS-Cluster while the RRHs' antennas in each cluster act as a single coherent antenna array distributed over a cluster region, and discuss its advantages. Moreover, we propose innovative solutions using C-RAN architecture to improve existing macro-diversity, mobility-management, and capacity-enhancement schemes. Finally, in Section 7, we draw the main conclusions and wrap up the article by discussing future work.

2. Shortcomings of today's RAN

In traditional cellular wireless networks, each BS is connected only to a fixed number of sector antennae and provides service to a small coverage area. However, such architecture is not compatible with today's users' requirements and presents several disadvantages. This section briefly presents the shortcomings that today's cellular networks are facing. In the next one we will emphasize how C-RAN together with novel provisioning and allocation methods has the potential to solve many – if not all – of these.

2.1. High power consumption

To offer broadband wireless network and increase the coverage, operators continually increase the number of BSs. This leads to a dramatic rise in power consumption and consequently translates into higher OPEX. Fig. 1(a) and (b) show the components of power consumption reported by China Mobile [1]; here, the majority of power is consumed at BSs of the RAN. In each BS, the RAN equipments only consume half of the power, while the other half is used by air conditioning and by other equipment.

2.2. Rapidly increasing CAPEX and OPEX

The proliferation of personal mobile computing devices 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 over the last few years. To satisfy such shift in consumer data-rate usage, mobile operators need to increase their network capacity. However, additional deployment and maintenance of a large number of stand-alone cellular BSs to meet the growing capacity demand are highly inefficient due to excessive capital and operating expenditures. Practically, up to 80% CAPEX of a mobile operator is spent on the RAN, which means that most of the CAPEX is spent on building up BSs. On the other hand, OPEX includes the

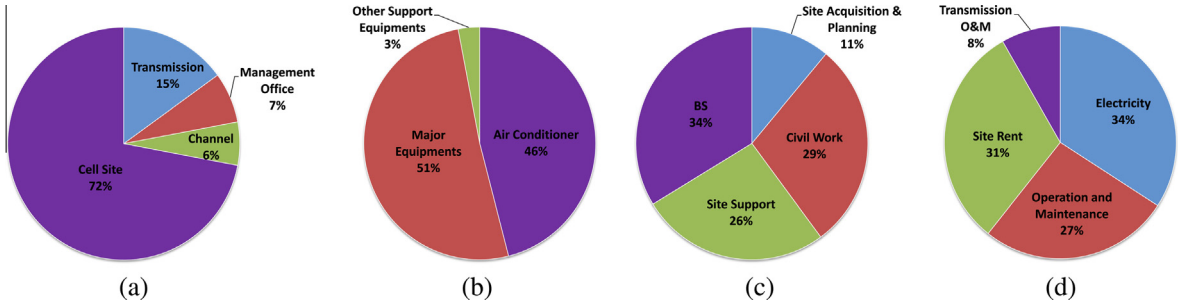


Fig. 1. Power consumption of (a) the Radio Access Network (RAN) and (b) the Base Stations (BSs); (c) CAPEX and (d) OPEX over 7 years.

costs for site and transmission network rentals, operation/maintenance, and bills from the power suppliers. Given a 7-year depreciation period for BS equipment, as depicted in Fig. 1(c) and (d), an analysis of the Total Cost of Ownership (TCO) shows that OPEX accounts for over 60% of the TCO, while the CAPEX only accounts for about 40%. Hence, the OPEX is a key factor in building future RANs.

2.3. Multi-standard environment

Today, BSs in wireless access networks make use of proprietary hardware designs and support specific standards. When the wireless network is upgraded, almost all of the network equipment must be replaced. Furthermore, during the transition, in order to satisfy the coexistence of new standards (such as WCDMA in 3G) and old standards (such as GSM in 2G), mobile operators must keep the old network and create another one for the new standard. Therefore, wireless network upgrades require huge financial investments and have often limited adoption of the emerging wireless technologies and algorithms.

2.4. Limited inter-BS cooperation:

Traditional cellular wireless systems are suffering from limited inter-BS data exchange and do not allow to fully exploit the potential of cooperative communication schemes like macro-diversity and collaborative spatial multiplexing. In general, message between the BSs need to be exchanged through the expensive backhaul links, and perhaps even over one-level higher in the aggregation hierarchy. Currently, to perform cooperative communication schemes, it is proposed to divide a set of neighboring cells into clusters and connect the BSs via the Back-haul Processing Unit (BPU). However, even in this case, exchanging data between BSs in different clusters requires traveling over backhaul links. Hence, the cost, latency, and scarce interconnect capacity among BSs have limited BS cooperation schemes in practice.

2.5. Explosive network capacity need

Global mobile traffic has been increased 66-fold with a Compound Annual Growth Rate (CAGR) of 131% between 2008 and 2013 [2]. On the other hand, the peak data rate has been only increased with a CAGR of 55% from UMTS to LTE-A, leading to a large gap between the CAGR of new air interface and the CAGR of customers' need. To fill this gap, new network architecture and infrastructure technologies need to be developed to further improve cellular-system performance.

2.6. Dynamic network load and low BS utilization

The number of active users at different locations varies depending on the time of the day. For example, during the day, the BSs in downtown office areas are the busiest, while at night, or in general during non-working hours, the BSs in residential or entertainment areas are the busiest. This movement of mobile network load based on the time of the day and the week is referred to as the "tidal effect". Today, each BS's processing capability is only used by the active users in its cell range, causing idle BSs in some areas/times and oversubscribed BSs in other areas. Static resource provisioning for the peak (worst case) at each cell site leads to grossly underutilized BSs in some areas/times while provisioning for the average leads to oversubscribed BSs in some areas/times.

3. C-RAN architecture

C-RAN is an evolution of the distributed cellular network where the BS computational resources are pooled in a central location. The main characteristics of C-RAN are: (i) centralized management of computing resources, (ii) reconfigurability of spectrum resources, (iii) collaborative communications, and (iv) 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 remote Virtual Base Stations (VBSs) housed in centralized BS pools, (2) the Base Band Unit (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. As a precautionary measure and to be on the safe side, the optical fiber transmission latency is limited to less than 1% of the PHY processing latency [6]. Hence, the range of VBS pool is

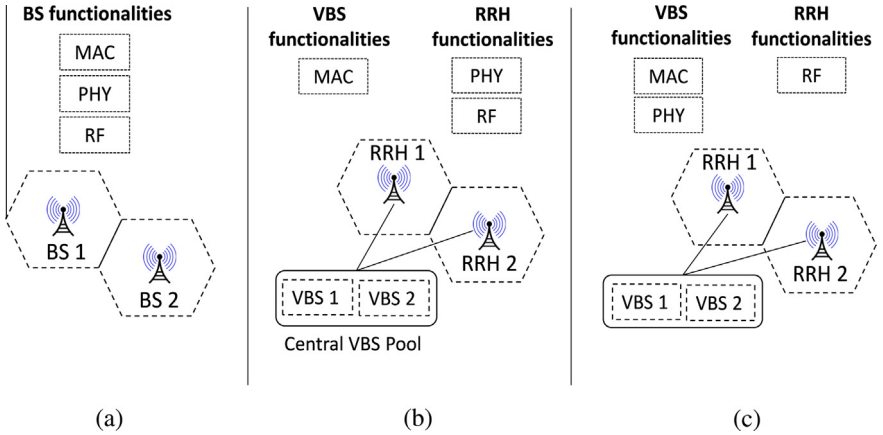


Fig. 2. Traditional Base Station (BS) architecture vs. two C-RAN architectures. (a) Distributed Base Station (BS) architecture; (b) partially-centralized architecture: only MAC processing is centralized in Virtual Base Station (VBS) pool; (c) fully-centralized architecture: VBS pool takes care of PHY and MAC processing.

limited by latency constraints of wireless system and services.

Based on the functionality of the RRH and VBS pool, two architecture have been suggested for C-RAN: partially- and fully-centralized architectures (see Fig. 2). In the "partially-centralized" (Fig. 2 (b)) architecture, the PHY processing is integrated into the RRH, while a VBS only takes care of MAC processing. This leads to the advantage of a lower volume of data to be exchanged between RRH and BBU (1/20–1/50 of the original baseband I/Q sample data [1]) and also the wireless resources can be scheduled on a global level. However, the capability of PHY cooperative techniques becomes lower and we still require remote equipment rooms in cell sites. In the "fully-centralized" (Fig. 2(c)), a RRH only takes care of Radio Frequency (RF) functionalities, while a BBU (VBS pool) takes care of both PHY and MAC processing. With a fully-centralized architecture, we are more capable to do cooperative techniques; however, such architecture requires a higher bandwidth to exchange data between RRH and BBU. Besides these advantages, C-RAN brings other benefits, which are listed below.

3.1. Lower power consumption

Since in C-RAN a group of BSs are centralized in a common place, the number of cell sites can be reduced several folds. Hence, the air conditioning and power consumption of other site support equipments can be dramatically reduced. In addition, since the cooperative interference reduction techniques can be applied among the RRHs, a higher density of RRHs is allowed. Hence, smaller cells with lower transmission power can be deployed, thus aiming for higher frequency reuse and capacity, while the network coverage is not affected. Deploying small cells reduces the energy used for signal transmission, which is especially helpful to reduce the RAN power consumption and increase MSs' battery time. Lastly, because the BBU pool is an aggregated collective resource shared among a large number of virtual BSs, a much higher utilization rate of processing resources and lower power consumption can be achieved via *statistical computing multiplexing*.

3.2. Lower CAPEX and OPEX

Since in C-RAN all the BBUs and site support equipments of a large region are co-located in a common datacenter, it is much easier and cost efficient for centralized management, operation, and maintenance compared to traditional RAN. In addition, the functionalities of the RRHs in the C-RAN architecture are much simpler, leading both their size and power consumption to be reduced so that they can be installed on top of buildings with minimum site support and management. Moreover, the RRHs only need the installation of the auxiliary antenna feeder systems, which allows the operators to speed the network construction up. Thus, operators can get large cost savings on site rental, operation, and maintenance, leading to lower OPEX and CAPEX.

3.3. Flexibility to add new standards

In C-RAN, the large scale BBU pool with high-speed low-latency interconnection, the common platform of Digital Signal Processor (DSP)/General Purpose Processors (GPP), and open Software Defined Radio (SDR) solution enable a cost-effective realization of VBSs. Therefore, in order to add/support new standards, there is no need to replace the equipment; conversely, it would suffice assigning new VBSs in the platform. As a result, CAPEX and OPEX costs associated with the wireless network upgrading can be eliminated altogether.

3.4. High speed inter-BS coordination

With the consolidation of BSs in a centralized VBS pool, such consolidated/co-located BSs can talk to each other at Gbps speeds and can communicate at low latencies, quasi real time. High-speed communication between the BSs can bring an extra degree of freedom to make optimal

decisions and fully exploit the potentials of cooperative techniques. As an example, a few approaches where cooperation among BSs can be beneficial are: (i) joint flow scheduling and load balancing, (ii) interference management, (iii) cooperative spatial multiplexing and macro-diversity, and (iv) mobility management.

3.5. Capacity improvement

In C-RAN, VBSs are able to exchange the signaling, traffic data, and CSI of active MSs in the system with low latency. This way, it becomes much easier to implement joint processing and scheduling algorithms so to mitigate ICI and improve spectral efficiency. For example, CoMP can efficiently be implemented under the C-RAN architecture.

3.6. High BS utilization rate

C-RAN is also suitable to handle non-uniformly distributed traffic due to its intrinsic load-balancing capability in the centralized BBU pool. Although the serving RRH changes dynamically according to the movement of the MSs, the serving BBU is still in the same BBU pool. As the coverage of a BBU pool is larger than in traditional BS, non-uniformly distributed traffic generated from MSs can be distributed in a VBS as this sits in the same BBU pool.

4. Software Defined Virtual Base Station Pool

Today's BSs are equipped with a set of heterogenous processing devices, each of which executes a specific task as defined at the design time. At the time of upgrading the network, almost all of the network equipment must be replaced. With DSP, GPP, and emerging SDR frameworks, we are now able to reconfigure the radio equipment. Large-scale BBUs endowed with high-speed, low-latency interconnection, plus the programmable DSP/GPP and open SDR solutions set the base for a VBS. In the C-RAN architecture a bunch of VBSs are pooled in a common BBU where a large amount of computing resources is available. Hence, VBS pool contains all the required processing resources of traditional BSs including entire digital signal processing at the PHY layer and packet processing at the MAC layer.

With virtualization technology we can dynamically allocate processing resources within a BBU to different VBSs. Whenever a user requests a service, computing resources need to be allocated for the corresponding service. This leads to a greater utilization of the processing resources and the ability to adjust in response to the tidal effects in different areas so to accommodate fluctuating demands. However, in general we are not able to pool all the VBSs together as there are some constraints to take into account. The range of VBS pool (BBU) is limited and depends on the latency constraints of the wireless networks. In C-RAN, the optical fiber transmission latency is suggested to be less than 1% of the PHY processing latency [6]. Assuming a PHY processing latency of 10 ms, the fiber transmission latency should be less than 0.1 ms. Since the signal speed through the fiber is $\approx 2 \times 10^8$ m/s, a signal path of 20 km has a latency of ≈ 0.1 ms. Consequently, a region with radius of 10 km is able to cover 314 km^2 of a metropolitan area, which may serve millions of users.

4.1. Technical challenges

BSs have strict real-time, low-latency, and high-performance requirements, to meet which the traditional virtualization technique is challenged. Specifically, to deploy real-time VBS pool the following requirements need to be met [1]:

- Advanced processing algorithms for real-time signals.
- High-performance, low-power processing for wireless signals.
- High-bandwidth, low-latency, low-cost BBU interconnection topology among physical processing resources in the baseband pool. These include the interconnection among the chips in a BBU, among the BBUs in a physical rack, and across multiple racks in datacenter.
- Efficient and flexible real-time operating systems to achieve virtualization of hardware processing resources management and dynamic allocation of physical processing resources to each VBS so to ensure processing latency and jitter control hardware-level support on virtualization.

There are only a few works that have started to address some of these challenges. In [5], the authors recommend that timing and synchronization system should have two parts: the first, namely, master time server, provides the accurate timing reference, while the second distributes the precise timing signal throughout the VBS pool and RRHs. The authors also suggest to use standardized interface technologies widely used in IT infrastructure (GbE, 10-GbE, InfiniBand, and PCIe) to interconnect BBUs. For hardware efficiency and flexible collaboration, the same authors also propose to separate the PHY and MAC layers into different platforms. In [6], a hierarchical management is suggested, where computing resources are assigned on demand and in real time to different radio operators, cells, or services. The authors of [6] also discuss the complexity of some resource-management algorithms and introduce different management schemes in simulated VBS pool. In [7], the constraints of PHY and MAC layers are analyzed and the VBS performance is optimized to meet the stringent real-time requirements of jitter and latency. The authors of [7] also present the first working prototype of a VBS pool on a multi-core IT platform; specifically, they show that their VBS pool prototype for WiMax can meet system requirements including synchronization, latency, and jitter. The authors of [8] propose some low-complexity algorithms to minimize the network power consumption of C-RAN, including the transport network and radio access network power consumption. They formulate the network power consumption and propose an algorithm to switch off one RRH at each step. Then, to reduce the complexity, they propose a three-stage group sparse beamforming framework. In [9], a partitioning and scheduling

framework is proposed which is able to reduce the compute resources by 19%. In [10], the authors present a flexible framework for small cells, called Fluidnet, which dynamically reconfigures the front-haul based on network feedback to maximize the amount of traffic demand and optimize the compute resource usage in the BBU pool. The authors of [11] consider the coordinated transmission problem to minimize the downlink power in C-RAN. In order to serve each MS, they determine a set of RRHs and the precoding vectors for the RRHs to minimize the total transmission power subject to the constraints on fronthaul capacity. In [12], the authors consider the C-RAN with finite-capacity backhaul links and propose a hybrid compression and message sharing strategy for downlink transmission to optimize the backhaul capacity utilization.

4.2. Computational requirement of a VBS

To allocate processing resources to the VBSs, we need to study the computational complexity of the PHY- and MAC-layer functionalities. Compared to the PHY layer, MAC layer requires less than 10% of the computation resources of whole BS, while PHY layer occupies 90% of the resources. Assuming there is serial processing, Table 1 shows the required instructions for the PHY layer in some typical wireless standards [5]. In the following, we study PHY and MAC complexity in detail.

4.2.1. Complexity of PHY functionality

Packet Error Rate (PER) of a receiver is primarily dictated by the equalization technique employed at the PHY layer. However, the computational complexity (execution time) and memory footprint of advanced signal processing and decoding algorithms (for equalization) that guarantee very low PER are very high. For example, the Viterbi algorithm was proposed to solve the maximum likelihood sequence detection problem [15] so to eliminate inter-symbol interference and exploit frequency diversity in wideband communication systems. The complexity of the Viterbi algorithm is $\mathcal{O}(n \cdot M^L)$, where n is the block size in terms of number of modulated symbols, M is the constellation size of the modulation scheme used (e.g., $M = 2$ for BPSK, $M = 16$ for 16-QAM), and $L = T_D/T_S$ represents the number of taps in the wireless channel model and is computed as the ratio of the delay spread of the wireless channel (T_D[s]) over the symbol duration (T_S[s]). The physical memory requirements of the VBS is of the order of M^L, which represents the size of the state space in the Viterbi algorithm (i.e., the memory of the wireless channel). The complexity and physical memory requirements of the Viterbi algorithm increases with the increase in the constellation size and, more importantly, with the increase in L. L is typically high in environments with a large number of signal paths (e.g., indoor or downtown setting), when a higher level modulation (with small symbol time) is used, and also when the bandwidth of operation is wide.

Table 1
Millions-Instruction-Per-Second (MIPS) requirements for key wireless standards on BSs.

	GSM	W-CDMA	WiMAX
User data throughput	14.4 Kb/s [13]	2 Mb/s [13]	20 Mb/s [14]
MIPS	100	6000	30,000

The workload of the PHY layer can be analyzed using the algorithm- or system-level behaviors. Fig. 3 shows the algorithm-level behavior of the PHY layer for an OFDM system. Based on the techniques used in each block and their complexity, we would need to provision the VBSs. The provisioning of the VBSs should be done in such a way that the required computational resources be available for processing the OFDM frames before a deadline. Usually, there is a strict time duration frame such as 2 or 3 ms. This means that the downlink or uplink processing of a frame should be finished within the duration of the frame itself. Generally, each cell consists of a number of sectors (N), and each of them has downlink and uplink logics. Hence, the PHY workload in each BS can be divided into $2N$ logics. Since there is no data dependency among the sectors, these logics can be executed in parallel on a platform. It has been reported that because of multicore processors and parallel processing, one IBM QS21 blade with two cell/B.E. processors can support computation corresponding to 60 Mb/s data throughput for both uplink and downlink at same time [16,5].

4.2.2. Complexity of MAC functionality

A simplified high-level view of typical MAC-layer functionalities include multiplexing protocols transmitted over the MAC layer (when transmitting) and decoding them (when receiving), user scheduling, fragmentation (when transmitting) of MAC Service Data Units (MSDUs) from the higher-layer protocol in order to create MAC PDUs (MPDUs) or aggregation (when receiving) of MPDUs to form MSDUs. MAC frame processing, irrespective of whether it is performed per mobile user (as in 3G cellular systems) or per frequency-time block (as in LTE or WiMAX), generally has a stringent time constraint (e.g., a frame needs to be prepared and transmitted every 5 ms in WiMAX MAC layer downlink) [17]. The *frame processing latency increases with the increase in the frame size*, which is dictated by the number and size of packets from the higher layers. This in turn is determined by the number and combination of different types of user data traffic. This observation is backed up by the authors in [17] who tried to profile a software implementation of the WiMAX MAC layer. Knowledge of typical processing latencies under different loads enables judicious provisioning as well as allocation decisions.

In [7] the performance of PHY and MAC under a real VBS-pool prototype are evaluated. The configuration consists of an IBM x3650 server with two Intel Clovertown processors 2.66 GHz (8 physical cores in total). The single MAC performance for 20 MHz bandwidth and 64-QAM modulation was measured, and the authors concluded that the configuration is able to support a payload throughput of over 30 Mbps. Madhavan et al. [17] compare the computational requirements of (a) conventional distributed network architecture with that of (b) a network where the MAC-layer functionalities of multiple VBSs are

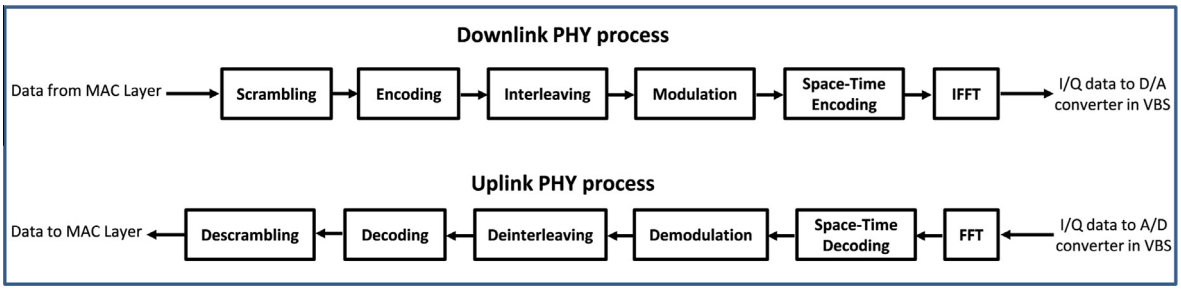


Fig. 3. Algorithm-level behavior of PHY layer (downlink and uplink) in a MIMO OFDM wireless BS.

consolidated in a common place while the PHY-layer processing stays close to the antenna site. It has been shown that the hardware requirement of the MAC consolidated network is much lower than that of a distributed one; in fact, the cloud computational requirement can be as small as half of a distributed network. This is because of the fact that packet traffic from the MS and hence the load offered to a BS is naturally random. In addition, different cells have different user arrival pattern based on their location leading different BSs have different peak traffic time. This intimates that the peak of the cumulative load offered to the cloud platform by multiple BSs will be smaller than the sum of their individual peak loads in the distributed setting. Consequently, lower computational resources is needed for centralized structure. The authors also show that when the number of consolidated BSs doubles, the multiplexing gain increases by more than double. So we can conclude that the larger the cloud provider, the better the obtainable infrastructure gain. However, a large cloud requires a larger investment for deploying high-speed links from RRH to the VBS pool. So, the optimum size of the cloud is limited by these cost factors rather the cost of the computing infrastructure, which only reduces for a bigger datacenter.

5. Demand-aware dynamic virtual base station

The number of active users at different localities varies depending on the time of the day as shown in Fig. 4. During the day, the BSs in downtown office areas (such as VBS #2 in the figure) are the busiest, while at night or non-working hours the base stations in residential or entertainment (such as VBS #1 and VBS #3) areas are the busiest. This movement of mobile network load based on the time of the day and the week is referred to as "tidal effect". Today, each BSs' processing capability is only used by the active users in its cell range. Static resource provisioning for the peak (worst case) at each cell site leads to grossly underutilized BSs in some areas/times, while provisioning for the average leads to oversubscribed BSs in some areas/times. We advocate *demand-aware resource provisioning* in which the VBSs will be dynamically resized in order to meet the fluctuating demands of the cellular network. As shown in Fig. 4(a), during working hours, VBS #2 will be provisioned with more computing resources compared to the ones serving a residential area (VBS #3) or a stadium (VBS #1). However, during the night, the VBSs serving the stadium (on a game night) or the residential

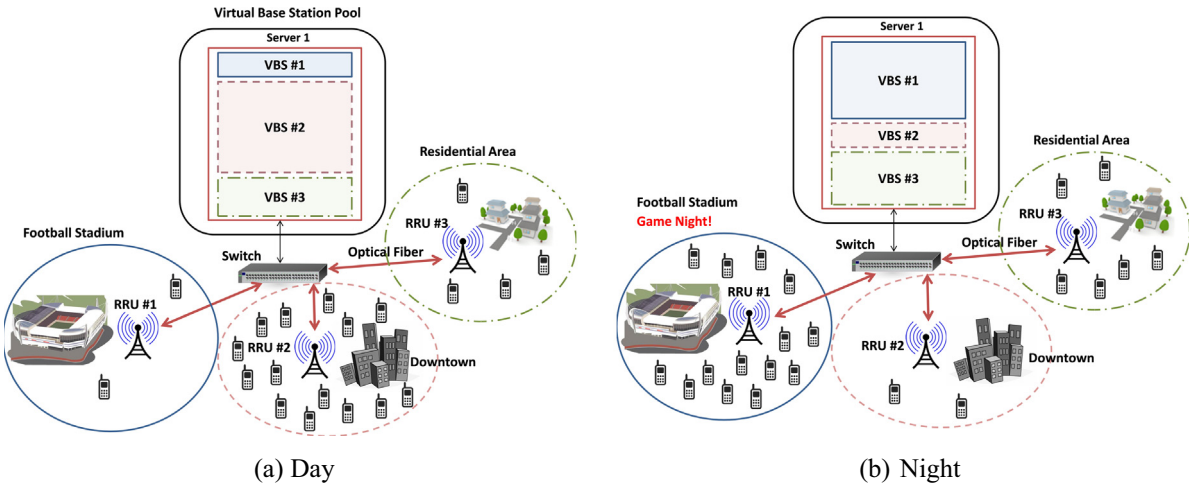


Fig. 4. The use of virtualization in C-RAN allows dynamic re-provisioning of spectral and computing resources (visualized here using different sizes) to Virtual Base Stations (VBSs) based on demand fluctuation; (a) and (b) illustrate the movement of mobile network load from the downtown office area to the residential and recreational areas over the course of a day.

area will be provisioned with more resources than the ones in downtown to meet the change in demand, as in Fig. 4(b).

5.1. Size of a VM

All or part of the communication functionalities of a VBS (such PHY, MAC, packet processing) are implemented in a Virtual Machine (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/s], memory and storage capacity [Bytes], and network interface speed [bit/s]. It primarily depends on the number of mobile users and types 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 of the scheduling and frame processing algorithms at the MAC layer. In addition, the complexity of communication optimization algorithms (such as the ones for inter-cell interference mitigation among densely deployed BSs in high-capacity RANs) may also dictate 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 is less than the frame deadline.

Our solution for dynamic provisioning (or re-provisioning) 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 anticipated and the computational resources are provisioned in advance for a limited 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 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. Even though the proactive component allows for a smooth transition and greater optimization (w.r.t. energy expenditure and resource utilization) in ensuing VM-allocation procedure, it cannot handle 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. The reactive component monitors the CPU/memory/network utilization of the VMs and triggers over- or under-provisioning alerts when there is a 'significant' mismatch between the expected resource utilization (based on the profile) and the actual observation.

5.2. Dynamic resource provisioning

To demonstrate the multiplexing gain (increase in the region of feasibility) that can be achieved through dynamic resource provisioning, we simulated on MATLAB R2013b the following scenario: two BSs, one serving indoor users (or users in the downtown area with a large number of obstacles) and another serving outdoor users (say a recreational area in the suburb). This simulation also serves to demonstrate the benefit of C-RAN over the traditional distributed BS architecture for cellular networks. As mentioned earlier, the computational complexity and the memory footprint of PHY- as well as MAC-layer functionalities depend on a number of factors. At each BS, we assume that each user's traffic belongs to one of the three following types, namely, (i) voice over IP (very low and constant bit rate), (ii) light browsing (bursty but low data rate), and (iii) streaming or downloading (high data rate), with equal probability and in the increasing order of priority. 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 the region of feasibility to be the total number of active users served by the BS pool with an acceptable blocking probability of 5%, which is a metric typically used in the context of voice calls. Here, for simplicity, we reuse the term to also convey acceptable level of service degradation in data traffic. Fig. 5 shows that dynamic resource provisioning (Case 3) increases the

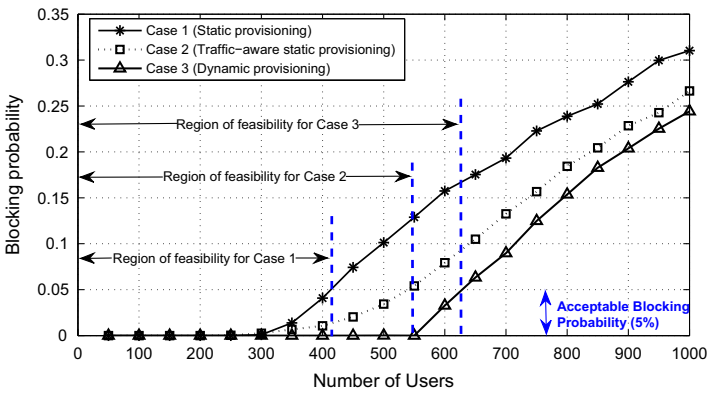


Fig. 5. Benefit of dynamic computing-resource provisioning for the VBSs at the remote datacenter: increase in the region of feasibility (w.r.t. active users).

region of feasibility (in terms of number of users) by as much as 50% compared to the simplest static provisioning case (Case 1). Knowledge of relative distribution of users among BSs can help improve the feasibility region (Case 2), but may result in chronic over- and/or under-provisioning when the demand fluctuation is very high. Greater benefits can be obtained when the distribution of users of different traffic types is unequal.

5.3. Quality of Service (QoS)-aware resource allocation

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

5.3.1. Consolidation of VMs

We advocate thermal-aware *VM consolidation* [18] for the VM allocation problem. Thermal awareness, which is the knowledge of *heat generation* and *heat extraction* at different regions inside a datacenter, 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 servers after VM consolidation; (2) the utilization of servers that are in the "better cooled" areas of the datacenter (i.e., with high heat extraction) can be maximized; (3) according to thermodynamics, heat can be extracted more efficiently (i.e., by doing a lower amount of work) by the cooling system from the consolidated server racks, which are hotter than non-consolidated server racks. In addition, consolidation on servers hosting VBSs of physically co-located RRHs allows for an efficient implementation of common functionalities such as signaling, CSI estimation for active users in a RAN as well as for joint processing and scheduling techniques (like CoMP in 4G) for inter-cell interference mitigation.

Thermal and energy awareness alone, however, are insufficient for guaranteeing high VBS performance and for maximizing energy and resource-utilization efficiency. As multiple VMs share the same server resources (CPU, memory, 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 the VM allocation procedure, we advocate the need 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 VM-allocation algorithms, thus making them QoS aware.

5.3.2. Split-VBS architectures

To improve user QoS and resource utilization in C-RAN, we can deploy different architectures for VBSs. Fig. 6 shows three possible split-VBS architectures apart from 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.

5.3.2.1. 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 datacenter with heterogeneous servers, exemplified as Server 1 and 2 in Fig. 6(a), we can match the workload of BS-stack components with the capabilities of specific hardware.

5.3.2.2. Many-to-one. In general, communication among BSs can improve cellular-system performance by exploiting the global and shared nature of information so to make optimal decisions. For instance, in BS-cooperation schemes, significant control information needs to be exchanged among neighboring BSs. Yet, cost, latency, and

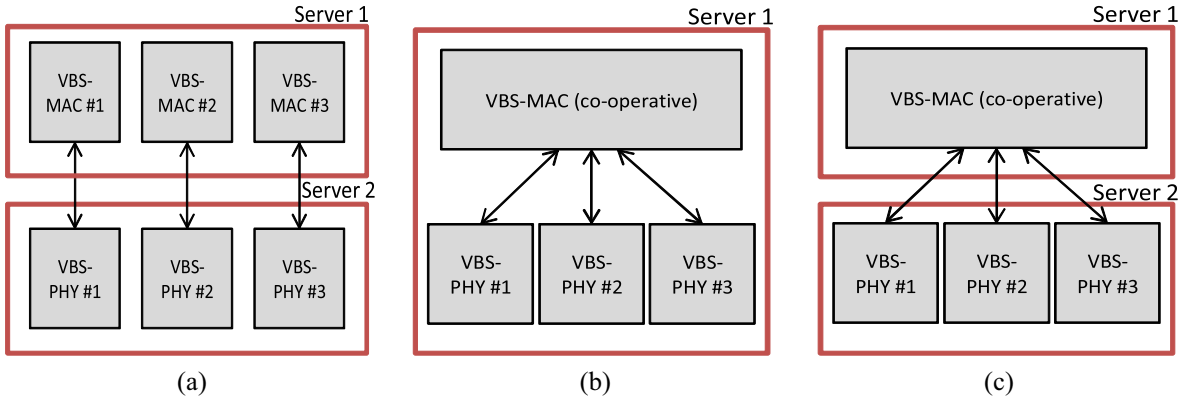


Fig. 6. Alternative split-VBS architectures: (a) one-to-one mapping between PHY and MAC (different servers); (b) many-PHY-to-one-MAC (one server); (c) many-PHY-to-one-MAC (different servers). Note that architectures (a) and (c) (multi-servers) exploit the heterogeneity in datacenter server hardware.

scarce interconnect capacity among BSs have been major impediments to the implementation of such schemes so far. We propose a split-VBS architecture, depicted in Fig. 6(b), in which the information of the MAC layers can be shared at Gbps speeds making very-low-latency inter-BS communication possible. As a result, faster mobility management, novel interference management, and advanced cooperative MIMO techniques can be implemented to improve the user QoS. Finally, in order to take advantage of the heterogeneous processing pool as well as of the high-speed inter-BS communication, we propose the architecture in Fig. 6(c).

6. Advantages of BS consolidation

In current distributed cellular systems, BSs can barely communicate with each other as messages among the BSs have to be exchanged through costly backhaul links. So the traditional distributed BS architecture is characterized by latency, cost, and scarce inter-BS communication capacity. In C-RAN, since all the VBSs are located in a common server, they can exchange data with each other at Gbps speeds. On the other hand, clustering the VBSs of the neighboring cells in the C-RAN architecture – together with enabling the coordination of the VBSs in the cluster – can improve the system performance by exploiting the extra degrees of freedom to make optimal decisions [19]. Here, we introduce the idea of *VBS-Cluster*, where (i) all the VBSs associated with a certain cluster are merged together and (ii) the RRHs' antennae in each cluster act as a single coherent antenna array distributed over the cluster region. Fig. 7 shows two VBS-Clusters, #1 (on the left) and #2 (on the right), where the sizes of the clusters

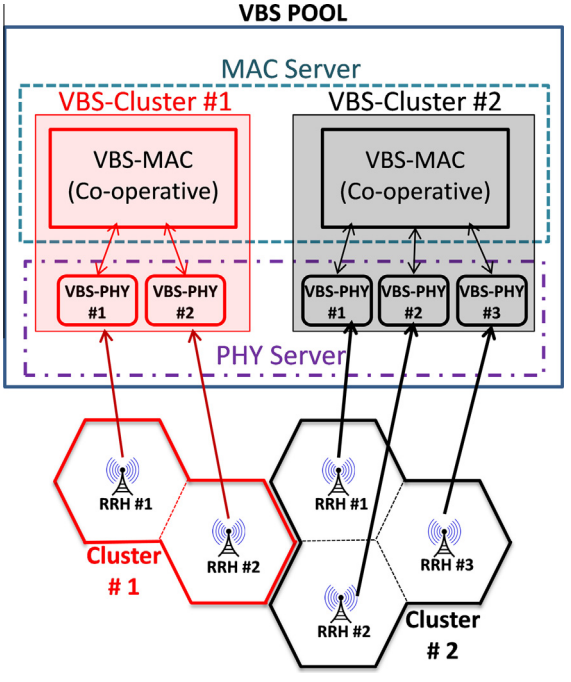


Fig. 7. Virtual Base Station Cluster (VBS-Cluster).

are 2 and 3, respectively. Since in the C-RAN architecture the 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. Moreover, we are able to assign each cell to different clusters in order for them to cooperate with each other using different techniques. Since associated VBSs of each cluster needs a high-data-rate communication to perform cooperative techniques, they have to be allocated in a same server in order to provide high speed inter-VBS connection. Moreover, as the number of active MSs in the cluster determines the size of VBS-Cluster, the resource allocation needs to be performed for each cluster. In the following we present a few scenarios where clustering the VBSs to enable cooperation improves the system performance.

6.1. Macro diversity schemes

One of the simplest and most effective technique to overcome the destructive effects of fading and co-channel interference is *diversity*. There are various types of diversity to combat the negative effects of wireless fading channels [15]. As shown in Fig. 8(a), for edge users (i.e., users in the dashed area) we can support macro diversity, i.e., we have access to a MS signal from different BS receivers. As multiple BSs receive the signal of a mobile user while it moves closer to the cell boundaries, *soft handoff* provides a form of receive diversity by considering different BSs as the different receive antennae. The optimal way to process the signals from these multiple antennae is *Maximal Ratio Combining (MRC)*, which combines the received signals constructively to boost the Signal-to-Noise Ratio (SNR). However, for combining the signals, we need to share the received signals rather than the decoded bit-stream; this huge overhead along with the scarce interconnect capacity among BSs renders MRC non-implementable in practice. Hence, in todays' cellular networks, the sub-optimal *selection combining* [15] is employed instead of MRC. To get an idea of the amount of data to be exchanged for MRC, let us consider a 5 ms WiMAX frame and a 10 MHz-channel bandwidth. The amount of data transferred is equal to: No. of samples * Bits per sample * No. of subcarriers * No. of OFDM symbols/Frame duration. The number of subcarriers here is 720 and the number of uplink symbols is 15. If 8 bits/sample were the signal resolution, then the traffic overhead requirement would total 52 Mbps. The packet decoding reliability is dependent on the signal resolution and the overhead requirement increases further if a reduced decoding error is desired. With scarce interconnect capacity among the BSs, this method is presently not implementable.

In C-RANs, when QoS-aware VBS allocation is employed, neighboring BSs will be co-located in the same physical server. As shown in Fig. 8(a), we divide each cell into 3 sectors and merge 3 neighboring sectors from different cells so to form a VBS-Cluster. It is clear that in this case each cell is associated with 3 different VBS-Clusters. Since, in each VBS-Cluster we have 3 different versions of the MS signal, we can apply MRC to improve the performance of the system. Fig. 8(b) shows the Bit Error Rate (BER) improvement (under fading channel) as the number of

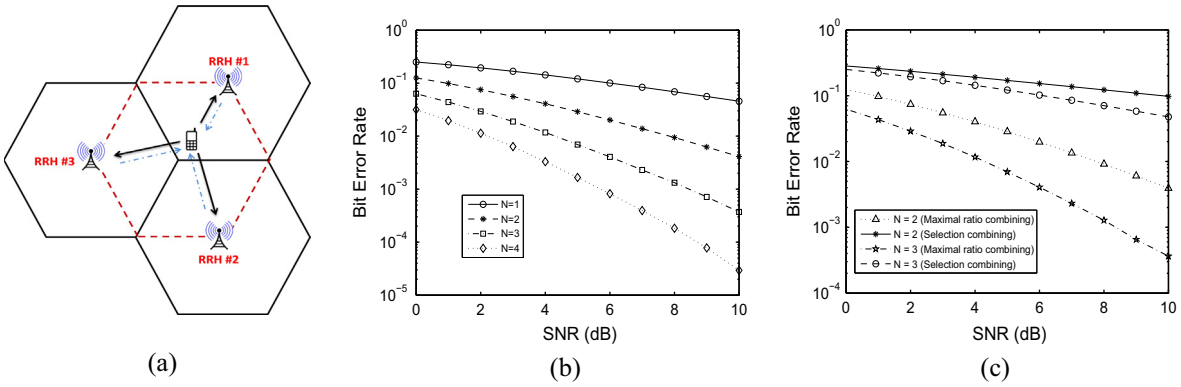


Fig. 8. Bit Error Rate (BER) improvement when Maximal Ratio Combining (MRC) is used to combine and decode the received signals at the different antennae; (a) macro diversity in C-RAN; (b) BER improvement as the number of receiving antennae N increases under fading channel; (c) BER improvement when MRC is used instead of the widely used selection combining [15].

receive antennae N used for MRC increases. Fig. 8(c) compares the performance of MRC (enabled by VBS co-location in C-RAN) with that of selection combining [15], which is employed in traditional distributed-BS cellular systems.

6.2. Mobility management

One trend to increase the spectral efficiency is to go for smaller cells; however, smaller cells lead to a higher number of handovers, especially in high mobility scenarios. In real-time streams like Voice over IP (VoIP) or Video on Demand (VoD), handover latency is a crucial factor to take into consideration. In general, handover sessions should be handled in such a way as to ensure minimum service disruption. Several handover schemes have been proposed, each shooting for a different tradeoff between spectrum resources and backhaul communication to reduce handover latencies [20].

The simplest handover scheme is Hard Hand-Over (HHO), in which the connection between the serving BS and MS is terminated before the connection between the new BS and the MS is started. For such scheme, Fig. 9 shows the sequence of messages that need to be exchanged between the MS, serving BS, and two target BSs. As it was studied in [21], the service disruption time caused by the exchange of these messages and the setup of the new connection between MS and new BS can be 250 ms or more, which is intolerable for some real-time services like VoIP. Interestingly, in [22], the authors studied how a co-located VBS pool eliminates the synchronization and ranging steps by sharing the overhead data between the serving and target VBS; VBS pooling also speeds authentication and Cell ID (CID) assignment up during the handover procedure. In addition to eliminating sync and ranging steps as well as speeding up the steps in handover procedure, C-RAN brings other improvements to handovers. Soft Hand-Over (SHO) is one of the Code Division Multiple Access (CDMA) handover schemes that can avoid service disruption as a MS can be actively connected to multiple BSs *simultaneously*. This contrasts with non-CDMA systems, in which a MS can *only* be connected to one BS at a time. In C-RAN architecture, since the VBSs are co-located in a common place and can communicate and exchange data and controlling signals with each other, we are able to actively connect a MS to multiple VBSs regardless of the modulation scheme. This means we are able to use SHO *both* for non-CDMA and CDMA schemes. By clustering the VBSs, as long as the MS is in a certain cluster, if the transmitted/received signal is weak to/from a MS then VBS-Cluster is able to change the serving RRH without any service disruption; in this case, a handover is needed less frequently, i.e., *only* when the MS wants/needs to change the VBS-Cluster.

To show the performance of our proposed solution in terms of number of handover sessions, we consider two mobility models: (1) Random Waypoint and (2) Gauss-Markov [23]. In the first model, a MS moves from its current location to a new one by choosing randomly a direction/angle d [rd] and a speed s [m/s] from pre-defined ranges, e.g., $[0, 2\pi]$ and $[s_{min}, s_{max}]$, respectively. After choosing these parameters, a MS moves to its new location by traveling for a certain time or distance. The model also includes pause time between changes in direction and speed. The second mobility model is designed to adapt to different levels of randomness by means of a tuning parameter: the direction and speed at the n^{th} step are calculated based regressively on those at the $(n-1)^{th}$ step and on a random variable (r.v.), as,

$$\begin{aligned} d_n &= \alpha d_{n-1} + (1-\alpha)\bar{d} + \sqrt{(1-\alpha^2)}d_{x_{n-1}} \\ s_n &= \alpha s_{n-1} + (1-\alpha)\bar{s} + \sqrt{(1-\alpha^2)}s_{x_{n-1}}, \end{aligned} \tag{1}$$

where d_n and s_n are the new direction and speed for the n^{th} step, α $(0 \leqslant \alpha \leqslant 1)$ is the tuning parameter to vary the randomness, $\bar{d}$ and $\bar{s}$ are constants representing the mean values of direction and speed, respectively, as $n \to \infty$, and $d_{x_{n-1}}$ and $s_{x_{n-1}}$ are r.v. from a Gaussian distribution. Table 2 represents the reduction in the number of handover sessions using VBS-Cluster. In the simulations, we performed a handover to the neighboring cell/cluster if both of the following conditions are met [24]: (1) If the signal strength from the neighboring cell/cluster exceeds that of the serving cell/cluster by a hysteresis (i.e., margin) level

Table 2
Reduction of no. of handovers using our VBS-Cluster strategy. [cell radius = 1 km, simulation area = 30×30 km^2, $s_{min} = 0$, $s_{max} = 30$ m/s, simulation time = 1 h, no. of MSs = 1000, $\bar{d} = \pi$, $\bar{s} = 15$ m/s, $d_{x_{n-1}} \sim \mathcal{N}(\pi, 1)$, $s_{x_{n-1}} \sim \mathcal{N}(15, 3)$, no. of simulations = 100].

Mobility model	Number of handovers			
	Without clustering	Cells/cluster = 3	Cells/cluster = 4	Cells/cluster = 5
Random waypoint	$5716 \pm 5\%$	$2318 \pm 4\%$	$1268 \pm 5\%$	$843 \pm 6\%$
Gauss-Markov	$3673 \pm 0.6\%$	$1682 \pm 1.1\%$	$711 \pm 2.3\%$	$457 \pm 2.9\%$

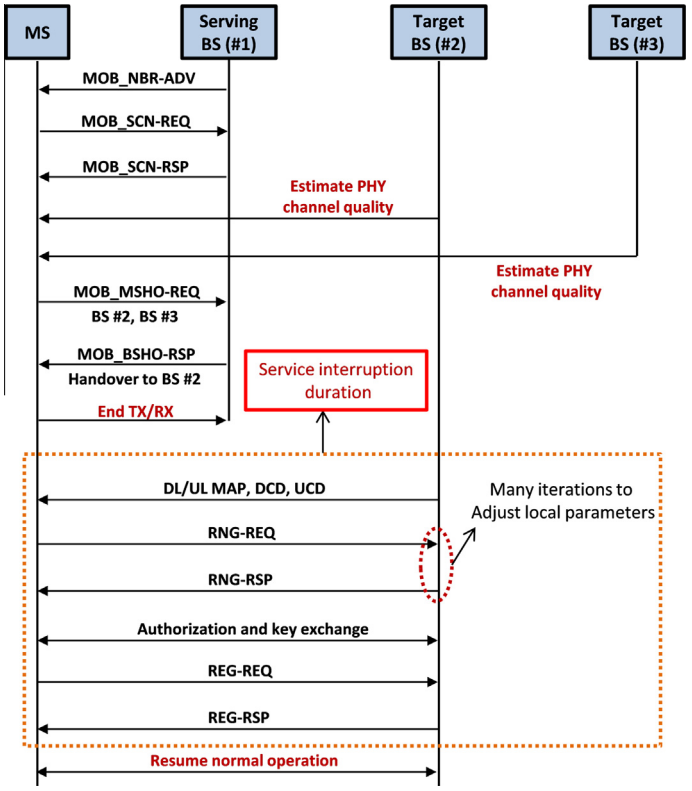


Fig. 9. Message exchanges among a MS and the BSs during a HHO.

of at least dB; and (2) If the distance from the serving cell/cluster exceeds that of the neighboring cell/cluster by more than 1.1 km. It is clear that the number of handovers decreases with the increase of the cluster size. However, by increasing the cluster size managing/canceling interference becomes more challenging.

6.3. Capacity enhancement

In conventional cellular network, each BS only sends/receives data to/from the MSs within its covered area, and uplink and downlink signals from the neighboring cells interfere with each other, which leads to low SINR and spectral efficiency for edge users. The work in [25] showed via simulations that as many as 30% of the users can be affected by cell-edge interference, in that they experience a SINR of 0 dB or less. In C-RAN, VBSs cooperate together and share traffic as well as signaling data so to improve the spectral efficiency.

The CoMP transmission and reception technique, which is based on cooperative MIMO, is one of the popular methods to mitigate the average interference and increase the spectral efficiency at the cost of increased receiver complexity [26,4,27]. In CoMP, a set of neighboring cells are divided into *clusters*; within each cluster, BSs are connected to each other via the Back-haul Processing Unit (BPU) and exchange Channel State Information (CSI) as well as MS signals. Coordination of the BSs within a cluster can decrease the ICI and improve the overall SINR. In the uplink, each BS receives a combination of MS signals from its own and from the other neighboring cells. Fig. 10(a) shows the uplink intra-cluster and inter-cluster interference. By combining the CSI from different cells and sharing the received signals at the BPU, CoMP is thus able to cancel the intra-cluster interference.

Since MSs use the omni-directional antennas to send the signal to the BSs, decreasing the interference in uplink is more challenging. Although CoMP is able to reject the

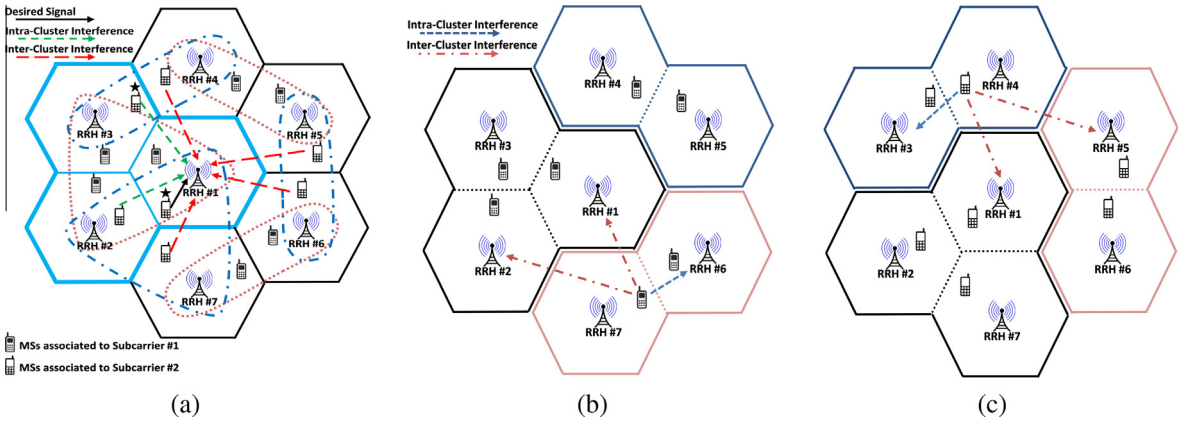


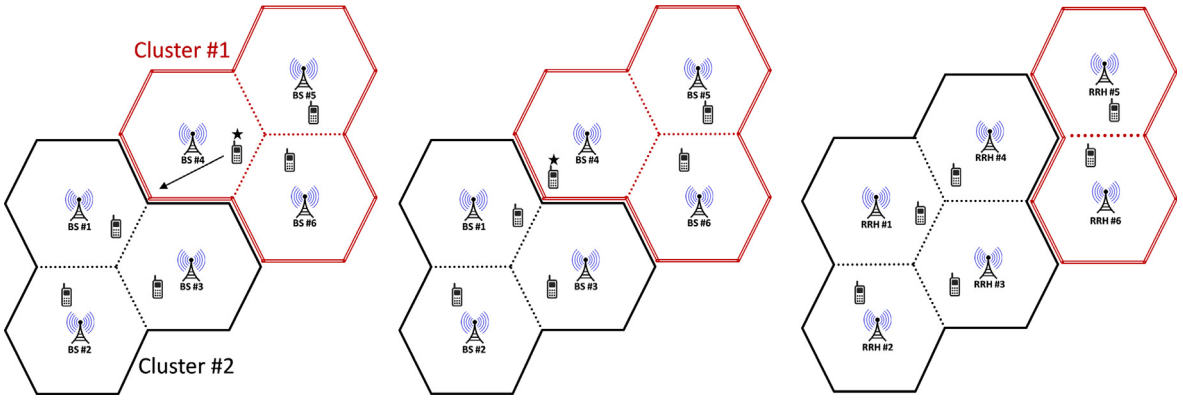
Fig. 10. (a) Uplink *intra-* and *inter-cluster* interference for subcarrier #2. The cluster (defined by bold blue lines) is omni-subcarrier and starred MSs have an intensive inter-cluster interference on the neighboring cells; in (b) and (c) the clusters are virtual uni-subcarrier and there is no cluster-edge MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

intra-cluster interference, it cannot mitigate the inter-cluster interference and cluster-edge MSs still suffer from this type of interference. Hence, in cellular network with frequency reuse factor equal to 1, the achieved system capacity is still significantly far from the interference-free capacity upper bound. Furthermore, one of the main requirements for Long Term Evolution (LTE) systems is the very low latency. In fact, the additional processing required for multiple-site reception and transmission as well as the communication among different BSs could add delay significantly and limit the cluster size. In addition to low-latency connections, BS clocks need to be in phase in order to enable proper operation of CoMP. This requires a highly accurate phase or time-of-day synchronization. To overcome these challenges, the BSs should be connected together in a form of a centralized Radio Access Network (RAN).

In traditional CoMP systems, however, only a fixed number of BSs can communicate with each other through the BPU and each cell is associated with a particular cluster. These limitations are avoided in C-RAN, where each cell can be associated with different clusters and different clusters can communicate with each other at very high speeds. Moreover, cluster boundaries are not set apriori, which means that we can add/remove cells to/from a cluster. These advantages allow us to form clusters and dynamically adjust their sizes based on the positions of MSs and RRHs in order to mitigate the inter-cluster interference. This means that in our solution the clusters are defined per subcarrier so that the defective impact of inter-cluster interference is low. In other words, unlike the traditional CoMP in which the clusters are "*omni-subcarrier*" (i.e., in each cluster all subcarriers are used), in our solution the clusters are "*uni-subcarrier*" (i.e., each cluster *only* deals with one subcarrier). Consequently, each cell may be involved in different clusters for different subcarriers.

To clarify the motivation, we use a network of 7-cell sites (as shown in Fig. 10) with two operating subcarriers. We also use different icons for MSs operating on different subcarriers. In Fig. 10(a), we assume that this network works under traditional CoMP and cells #1, #2, and #3 form an omni-subcarrier cluster (the cluster boundaries are shown with thick blue lines). In this case, internal cluster-edge MSs associated with subcarrier #2 (which are distinguished by a 'star') have a destructive inter-cluster interference on the neighboring external RRHs (RRH #4 and RRH #7). To address this problem, we propose to form *virtual uni-subcarrier clusters* based on the position of MSs and RRHs. In our solution, called Dynamic Joint Processing (DJP), virtual clustering must be done is such a way that the internal MSs have minimum inter-cluster interference on the neighboring virtual clusters. To do this, we need to measure the received power from each MS to the internal and external RRHs and decide to change the serving cluster if the interference on external RRH is greater than the interference on internal RRHs. In Fig. 10(a), dotted and dot-dash lines show uni-subcarrier clustering of cell site associated to subcarrier #1 and #2, respectively. Fig. 10(b) and (c) also show the uni-subcarrier clusters for subcarrier #1 and #2, respectively. For example, cells #1, #2, and #3 form a uni-subcarrier cluster for subcarrier #1 (Fig. 10)) and cells #1, #2, and #7 form a uni-subcarrier cluster for subcarrier #2 (Fig. 10(c)). Note that each cell is associated with two uni-subcarrier clusters, while in traditional CoMP each cell is only associated with one omni-subcarrier cluster. The clustering is done is such a way that there is no cluster-edge MSs and the received power from MSs to external RRHs is very low. This is because with uni-subcarrier clustering, all the internal MSs are as far as possible from external RRHs and situated in the center of the cluster.

Moreover, in our DJP solution, the cluster size of coordinated cell sites is not fixed. Since we operate under the C-RAN architecture and have access to all of the VBSs held in a common place, cluster size can be changed dynamically based on the MSs and the RRHs positions. We use Fig. 11 to clarify this property. In particular, Fig. 11(a) shows 2 omni-subcarrier clusters in CoMP. For a specific subcarrier in this setting, as long as all the associated MSs are far from the neighboring inter-cluster BSs, the performance of the CoMP system is acceptable and the



(a) Conventional CoMP (b) Conventional CoMP (c) Dynamic Clustering Under C-RAN Architecture

Fig. 11. Dynamic change of cluster size based on the MSs position: (a) there is no cluster-edge MS and no severe interference on the neighboring clusters; (b) the starred MS from cluster #1 is a cluster edge MS and has a high defective impact on its neighboring cluster; (c) because of the high defective impact of starred MS on cluster #2, cell #4 is removed from cluster #1 and added to cluster #2. In this case starred MS has not a severe defective impact on cluster #1 and its impact on cluster #2 can be canceled by the use of CoMP.

defective impact of the inter-cluster interference is low. However, as depicted in Fig. 11(b), if one of the MSs (starred MS in the figure) reaches the cluster edge, then the defective inter-cluster interference on BS #1 and #3 for the specific subcarrier would increase and the overall performance of cluster #2 would decrease. On the other hand, cluster-edge users have low defective impact on the intra-cluster BSs. For instance, in Fig. 11(b), the starred MS has low impact on the performance of BS #5 and #6, but high impact on BS #1 and #3. However, CoMP is not able to cancel the interference on BS #1 and #3, and is only able to cancel the interference on BS #5 and #6, which is not crucial. Since the cluster boundaries are fixed, we cannot do any high-speed cooperation with inter-cluster cells to decrease the inter-cluster interference. Fig. 11(c) shows the operation under C-RAN: in this case, since all of the VBSs are co-located in a common place, we are able to add/remove cells to/from a certain cluster; when a certain MS from a uni-subcarrier cluster changes its position potentially leading to a high defective impact on the neighboring uni-subcarrier cluster, we can remove the cell from the current cluster and add it to the neighboring cluster. As it is shown in Fig. 11(c), because of the high defective impact of starred MS on cluster #2, cell #4 was removed from cluster #1 and added to cluster #2.

In *DJP*, we consider that frequency band has a set of subcarriers $\mathcal{F}=\{f_1,\ldots,f_K\}$ (K is the total number of subcarriers), for each subcarrier the network has a set of virtual uni-subcarrier clusters $\mathcal{J}^k=\{1,\ldots,J^k\}$ $(1\leqslant k\leqslant K)$, each virtual cluster consists of a set of RRHs $\mathcal{M}_j^k=\left\{1,\ldots,M_j^k\right\}$ $(1\leqslant j\leqslant J^k)$, and in each virtual cluster there is a set of active MSs $\mathcal{N}_j^k=\left\{1,\ldots,N_j^k\right\}$ $\left(1\leqslant j\leqslant J^k\right)$. We measure the received power (in dB) from the MS $n_i^k\in\mathcal{N}_i^k$ by the RRH $m_l^k\in\mathcal{M}_l^k$ at time t as,

$$P_{rx}(n_i^k,m_l^k,t)=P_{tx}(n_i^k,t)-PL(n_i^k,m_l^k,t)-P_{\text{fading}}(n_i^k,m_l^k,t) \tag{2}$$

where $PL(n_i^k,m_l^k,t)$ is the large scale path loss between the MS n_i^k and the RRH m_l^k at time t, $P_{tx}(n_i^k)$ is the transmitted power of the MS n_i^k, and $P_{\text{fading}}(n_i^k,m_l^k,t)$ is the time-varying shadow fading loss. Since CoMP takes care about the intra-cluster interference, our goal is to minimize the inter-cluster interference. To do this, we measure the summation of received inter- and intra-cluster interference power from the MS n_i^k to the neighboring and serving clusters,

$$\begin{aligned}P_{ex}(n_i^k,j,t)&=\sum_{\forall m_j^k\in\mathcal{M}_j^k}P_{rx}\left(n_i^k,m_j^k,t\right),\\ P_{in}(n_i^k,i,t)&=\sum_{\forall m_i^k\in M_i^k,m_i^k\neq n_i^k}P_{rx}(n_i^k,m_i^k,t),\end{aligned} \tag{3}$$

where $P_{ex}(n_i^k,j,t)$ is the received inter-cluster interference from MS n_i^k by the jth virtual cluster and $P_{in}(n_i^k,i,t)$ is the received intra-cluster interference from MS n_i^k by the its serving virtual cluster (ith cluster). Then, we find the cluster which receives maximum inter-cluster interference from MS n_i^k and select it as the nominated cluster,

$$P_{j_{max}}(n_i^k,t)=\max_{\substack{1\leqslant j\leqslant J^k\\ j\neq i}}P_{ex}(n_i^k,j,t), \tag{4}$$

where $P_{j_{max}}(n_i^k,t)$ is the maximum inter-cluster interference from MS n_i^k and j_{max} is the index of nominated cluster to be added to the serving cell in the kth subcarrier at time t. In each iteration, we remove the serving cell from serving cluster and add it to the nominated cluster if $P_{j_{max}}(n_i^k,t)$ exceeds $P_{in}(n_i^k,i,t)$ by a hysteresis threshold *thr* (dB). To show the performance of our proposed solution, we compare the Cumulative Distribution Function (CDF) of the Signal to Interference Ratio (SIR) for different schemes (Fig. 12). To implement the conventional CoMP scheme, we consider the omni-subcarrier clusters of size 3. To compare DJP with conventional CoMP, we consider virtual uni-subcarrier clusters whose size ranges from 2 to 4 and set *thr* equal to 6.2 dB. In the simulation we compare DJP with

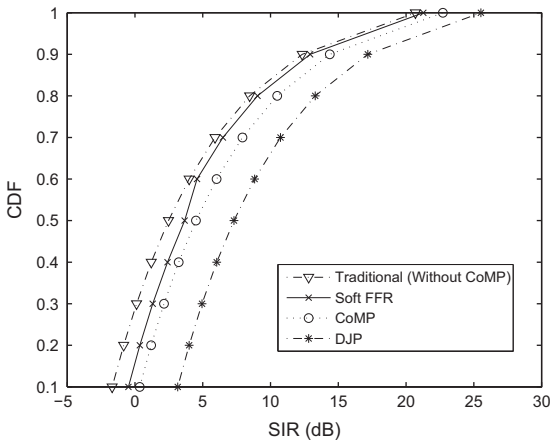


Fig. 12. Cumulative Density Function (CDF) vs. Signal-to-Interference Ratio (SIR) for different schemes.

traditional cellular network (without Inter-Cell Interference Coordination (ICIC)), Soft FFR [28], and Regular CoMP.

7. Conclusions and future work

We discussed provisioning and allocation methods to support Cloud Radio Access Network (C-RAN), in which all the BS computational resources are pooled at a central location, e.g., a set of physical servers in a datacenter. First, we presented the drawbacks of current cellular networks, which prevent from meeting today's users' data-rate requests. Then, we introduced C-RAN – a clean-slate design allowing for dynamic reconfiguration of computing and spectrum resources – and discussed about the functionalities that can be implemented in a fully- and partially-centralized architecture. We proposed novel provisioning and allocation methods of Virtual Base Stations (VBSs) in the Base Band Unit (BBU), and discussed their pros and cons. Last, we studied the complexity for provisioning VBSs: specifically, in order to overcome the mobile traffic load, a.k.a. tidal effect, and utilize the computational resources efficiently, we proposed proactive as well as reactive dynamic resource provisioning algorithms and investigated different Virtual Machine (VM) consolidation and allocation techniques. We also presented the idea of VBS-Cluster and discussed a number of solutions in which by clustering the VBSs we can improve the system performance.

8. Future work

To validate the proposed ideas on a real-time emulation, we are working on testbed implementation. There are three different implementations of software BS protocol stack, *OpenBTS* [29] and *OpenLTE* [30], which are open source, as well as *Amarisoft LTEENB* [31], which is a licensed software from Amarisoft SARL. OpenBTS is a software-based Global System for Mobile (GSM) communication access point, allowing standard GSM-compatible mobile phones to be used as Session Initiation Protocol (SIP) endpoints in VOIP networks. OpenLTE is an open source implementation of the 3GPP LTE specifications with focus on transmission and reception of the downlink. Last, Amarisoft LTEENB is a LTE BS software that has hundreds of tunable parameters at both the MAC and PHY layers. OpenBTS can be used to emulate centralized BS pools for GSM networks, whereas OpenLTE and Amarisoft LTEENB can be used to emulate centralized BS pools for LTE networks. At the Rutgers NSF Center for Cloud and Autonomic Computing (CAC), our state-of-the-art computing equipment and controllable CRAC system are representative of a small datacenter that can house the VBSs in C-RANs; the configuration consists of two Dell M1000E Modular Blade Enclosures, necessary interconnect/management infrastructure, and a supervisory node. Both transceiver ends of the platform are software radio USRP systems so to be able to implement the new structure/algorithm at both transmitter and receiver. For each VBS, we run Amari LTE 100 on 64 bits Linux on the available server and provision it with the required computing resources. We also use USRP N210, SBX RF board, and antennae for SISO and MIMO at the BS side. For the MS, LTE USB Modems are deployed.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. CNS-1319945.

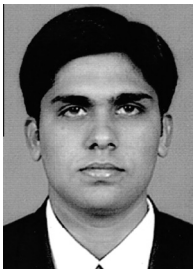
References

[1] C.M.R. Institute, "C-RAN: The Road Towards Green RAN," White Paper, ver 3, 2014.
[2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update," Cisco Public Information, Feb. 2013.
[3] D. Astély, E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, S. Parkvall, Lte: the evolution of mobile broadband, IEEE Commun. Magaz. 47 (4) (2009) 44–51.
[4] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, K. Sayana, Coordinated multipoint transmission and reception in lte-advanced: deployment scenarios and operational challenges, IEEE Commun. Magaz. 50 (2) (2012) 148–155.
[5] Y. Lin, L. Shao, Z. Zhu, Q. Wang, R. Sabhikhi, Wireless network cloud: architecture and system requirements, IBM J. Res. Develop. 54 (1) (2010) 1–4.
[6] V. Marojevic, I. Gomez, P. Gilabert, G. Montoro, A. Gelonch, Resource management implications and strategies for SDR clouds, Analog Integr. Circ. Signal Process. 73 (2) (2012) 473–482.
[7] Z. Zhu, P. Gupta, Q. Wang, S. Kalyanaraman, Y. Lin, H. Franke, S. Sarangi, "Virtual Base Station Pool: Towards a Wireless Network Cloud for Radio Access Networks," in: Proc. of the ACM Intl. Conf. on Computing Frontiers (CF), May 2011.
[8] Y. Shi, J. Zhang, K. Letaief, Group sparse beamforming for green cloud-ran, IEEE Trans. Wireless Commun. (2014).
[9] S. Bhaumik, S.P. Chandrabose, M.K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, T. Woo, Cloudiq: a framework for processing base stations in a data center, in: Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, ACM, 2012, pp. 125–136.
[10] K. Sundaresan, M.Y. Arslan, S. Singh, S. Rangarajan, S.V. Krishnamurthy, Fluidnet: a flexible cloud-based radio access network for small cells, in: Proceedings of the 19th Annual International Conference on Mobile Computing and Networking, ACM, 2013, pp. 99–110.
[11] V.N. Ha, L.B. Le, N. Dao, "Energy-efficient coordinated transmission for cloud-rans: Algorithm design and trade-off," in: 48th Annual Conference on Information Sciences and Systems (CISS), 2014, pp. 1–6.

[12] P. Patil, W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," Information Theory and Applications Workshop (ITA), 2014, pp. 1–6.
[13] M. Alsliety, D.N. Aloi, "Signal processing choices and challenges for SDR in telematics," in: 9th International Symposium on Signal Processing and Its Applications (ISSPA), Feb. 2007, pp. 1–4.
[14] Q. Wang, D. Fan, Y.H. Lin, J. Chen, Z. Zhu, "Design of BS Transceiver for IEEE 802.16 E OFDMA Mode," in: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2008, pp. 1513–1516.
[15] D. Tse, P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.
[16] L. Mitola III, Technical challenges in the globalization of software radio, IEEE Commun. Magaz. 37 (2) (1999) 84–89.
[17] M. Madhavan, P. Gupta, M. Chetlur, "On Quantifying Multiplexing Gains in a Wireless Network Cloud," in: IEEE International Conference on Communications (ICC), Jun. 2012, pp. 3212–3216.
[18] E. Lee, H. Viswanathan, D. Pompili, "VMAP: Proactive Thermal-aware Virtual Machine Allocation in HPC Cloud Datacenters," in: Proc. of IEEE Intl. Conf. on High-Performance Computing (HiPC), Dec. 2012.
[19] A. Hajisami, H. Viswanathan, D. Pompili, "Cocktail Party in the Cloud: Blind Source Separation for Co-operative Cellular Communication in Cloud RAN," in: IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS), 2014, pp. 37–45.
[20] I.L.S. Committee et al., "IEEE Standard for local and metropolitan area networks Part 16: air interface for fixed and mobile broadband wireless access systems amendment 2: physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1," in: IEEE Std 802.16-2004/Cor 1-2005, 2005.
[21] W. Jiao, P. Jiang, Y. Ma, "Fast handover scheme for real-time applications in mobile wimax," in: IEEE International Conference on Communications (ICC), 2007, pp. 6038–6042.
[22] P. Gupta, A. Vishwanath, S. Kalyanaraman, Y. Lin, "Unlocking Wireless Performance with Co-operation in Co-located Base Station Pools," in: Proc. of the Intl. Conf. on Communication Systems and Networks (COMSNETS), Jan. 2010.
[23] T. Camp, J. Boleng, V. Davies, A survey of mobility models for ad hoc network research, Wireless Commun. Mobile Comput. 2 (5) (2002) 483–502.
[24] K. Itoh, S. Watanabe, J. Shih, T. Sato, Performance of handoff algorithm based on distance and RSSI measurements, IEEE Trans. Vehic. Technol. 51 (6) (2002) 1460–1468.
[25] B. Ramamurthi, "Cutting edge at the cell edge: co-channel interference mitigation in emerging broadband wireless systems," in: IEEE International Communication Systems and Networks and Workshops (COMSNETS), 2009, pp. 1–7.
[26] Y.-H. Nam, L. Liu, Y. Wang, C. Zhang, J. Cho, J.-K. Han, "Cooperative communication technologies for lte-advanced," in: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010, pp. 5610–5613.
[27] J.G. Andrews, Interference cancellation for cellular systems: a contemporary overview, IEEE Wireless Commun. 12 (2) (2005) 19–29.
[28] T.D. Novlan, R.K. Ganti, A. Ghosh, J.G. Andrews, Analytical evaluation of fractional frequency reuse for OFDMA cellular networks, IEEE Trans. Wireless Commun. 10 (12) (2011) 4294–4305.
[29] OpenBTS: An Open Source Implementation of GSM Specification. <http://wush.net/trac/rangepublic>.
[30] OpenLTE: An Open Source Implementation of 3GPP LTE Specification. <http://sourceforge.net/p/openlte/home/Home/>.
[31] Amarisoft LTE Software Base Station. <http://www.amarisoft.com/?p=amarilte>.

Dario Pompili is an Assoc. Prof. in the Dept. of Electrical and Computer Engineering (ECE) at Rutgers University, NJ, where he is the director of the Cyber-Physical Systems Laboratory (CPS Lab) and the site co-director of the NSF Center for Cloud and Autonomic Computing (CAC). He received a Ph.D. in ECE from the Georgia Institute of Technology (GaTech) in 2007. In 2005, he was awarded GaTech BWN-Lab Researcher of the Year. He had previously received his 'Laurea' (integrated B.S./M.S.) and Doctorate degrees in Telecommunications and Systems Engineering from the University of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. In 2011, Dr. Pompili received the NSF CAREER award to design efficient communication solutions for underwater multimedia applications and the Rutgers/ECE Outstanding Young Researcher award. In 2012, he received the ONR Young Investigator Program (YIP) award to develop an uncertainty-aware autonomic mobile computing grid framework as well as the DARPA Young Faculty Award (YFA) to enable complex real-time information processing based on compute-intensive models for operational neuroscience. He is a Senior Member of both the IEEE Communications Society and the ACM.

Abolfazl Hajisami is a graduate student with the Dept. of ECE at Rutgers University, NJ, where he is pursuing a Ph.D. under the guidance of Dr. Dario Pompili at the CPS Lab. He received his M.Sc. and B.Sc. Degrees from Sharif University of Technology and from Shahid Beheshti University (Tehran, Iran), respectively. In his Master's thesis, he worked on the application of blind source separation in image watermarking. Currently, he is working on provisioning and allocation of virtual base stations in the Cloud Radio Access Network (C-RAN).



Hariharasudhan Viswanathan received his Ph.D. Degree in ECE from Rutgers University in June 2014, where he pursued research in the fields of mobile computing, datacenter management, and wireless networking in the CPS Lab under the guidance of Dr. Dario Pompili at the NSF CAC. Previously, he had received his B.S. in ECE from the PSG College of Technology, India and his M.S. in ECE from Rutgers University, in 2006 and 2009, respectively.