EAI Endorsed Transactions
on Wireless Spectrum

Research Article ICST.ORG

Survey on Cloud Radio Access Network

Reeta Chhatani [1,*] and Alice Cheeran [1]

[1] Department of Electrical Engineering, Veermata Jijabai Technological Institute, India.

Abstract

The existing wireless network will face the challenge of data tsunami in the near future. Densification of network will deal huge data traffic but will increase the interferences and network cost. At the same time, the existing wireless network is underutilized due to dynamic traffic. To deal with this adverse scenario, a change in the current network architecture is required. Based on virtualization, Cloud Radio Access Network (CRAN) was proposed for wireless network. In CRAN the functionality of base station will be distributed into base band unit (BBU) and remote radio heads (RRH) which will achieve benefits of centralization.

This paper presents a survey on CRAN centring on optimized resource allocation, energy efficiency and throughput maximization under fronthaul capacity. The existing solution and future opportunities in CRAN are also summarized.

Keywords: Cloud radio access network, Virtualization.

Received on 01 September 2015, accepted on 24 November 2015, published on 19 January 2016

doi: 10.4108/eai.19-1-2016.151001

1. Introduction

The mobile traffic analysis shows that globally the growth in mobile data traffic will grow 13-fold from 2012 to 2017, with a compound annual growth rate of 66% [1]. To support such huge traffic one of the solution is network densification, which will increase inter-channel interference as well as capital and operation expenditure (CAPEX and OPEX). Evolving mobile network standards will also provide a broadband capacity, so the coexistence of different standards needs to be supported by a telco operator that requires large investment and constant modifications and improvements of the network. At the same time the traditional network architecture remains underutilized as it is designed to handle peak traffic load which is present only for some time duration in a day. Also as the traffic is moving from one cell (urban area – business place) to another cell (town – home) during a day, this will cause great wastage of resources in low traffic area. This under-utilization is mainly due to confined architecture of existing network which do not allow sharing of radio resources over different network operators.

The existing network fails to efficiently utilize available resources, but have to deal with upcoming challenges of mobile data tsunami and coexistence of different standards. To handle this situation network architecture with substantial change has to be designed. Optimization is possible by centralization that incorporates coordination and intelligence in utilizing resources which reduces the investment. The centralization can be achieved through the concept of virtualization that allows sharing physical resources by separating them virtually. Radio access network (RAN) which is an essential part of wireless network can be modified to new architecture known as Cloud RAN (CRAN). CRAN can act as the architect to perform the sharing of resources in a centralized manner. CRAN will have a cloud of radio resources which are separated virtually and shared dynamically as per requirement [2].

Figure 1 shows CRAN architecture. It has two entities centralized base band units (BBUs) and localized remote radio heads (RRHs). BBUs are implemented on IT platform

[*]Corresponding author. Email: [1]reetagaokar@gmail.com; [2]ancheeran@vjti.org.in

that contains virtual base station (vBS) instances which can be configured in a centralized manner. RRHs are placed over a geographical area to provide wider coverage and capacity. RRH is the simple antenna system thus makes ease to speed up network footprint. BBU and RRH are communicating through high capacity low latency fronthaul.

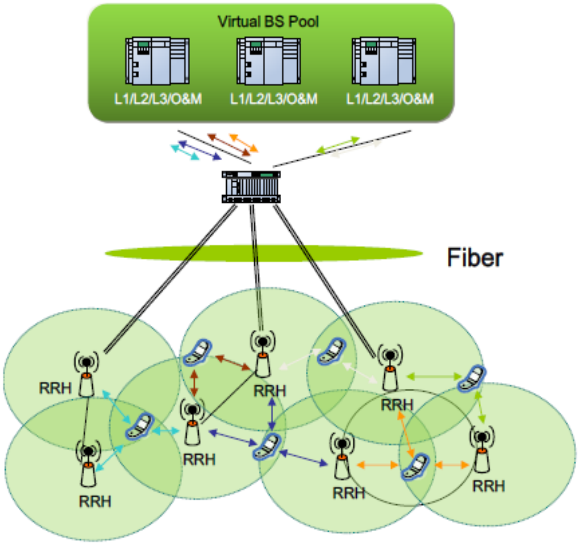

Figure 1. CRAN architecture [2]

This revolutionary centralized CRAN architecture provides the following benefits [2]:

1) Increase in capacity and reduction in inter-channel interferences: Centralized BBU is the pool of resources which can be shared dynamically and cooperatively among the multiple cells of different operators. Thus the resources can be utilized effectively as per service demand. The inter-channel interference can be reduced due joint scheduling and processing.
2) System throughput improvement: CRAN allows dense RRH deployment schemes in areas that express high throughput needs.
3) Reduction in capital and operational expenditure: As BBU is centralized and only RRHs are at different cell sites, the deployment and maintenance cost of per cell site has reduced greatly.
4) Coexistence of multiple standards: Centralized BBU can support multiple standards which can be effectively utilized as per user demands.
5) Green Radio: CRAN improves energy efficiency by reducing number of cell sites thereby reducing power consumption of site and equipments. Also during low traffic period, underutilized BBUs can be turned off and their traffic will be migrated to active BBUs.

The reminder of this paper is organized as follows: Section 2 describes resource allocation schemes in CRAN. The concept of green network is given in Section 3. In section 4, the constraint on cellular throughput due to CRAN architecture is explored. Section 5 gives the overview of the existing virtualized solutions and emerging research of CRAN architecture. Finally section 6 concludes this survey.

2. Resource Allocation and Management in CRAN

The virtualized CRAN resource allocation objective is to maximize the utility function which is defined in terms of the allocated rate, QoS or revenue generated by optimally utilizing resources. Maximization of utility function is the subject to achieving atleast minimum reserved bandwidth requested by service provider (SP) and the total achieved resources for all SPs must not exceed the resources available in the network.

Network virtualization substrate (NVS) algorithm [3] achieves this objective through flow level virtualization. Figure 2 shows the NVS algorithm which is implemented as MAC procedure on PicoChip WIMAX base station. The prototype consists of a WiMAX Profile-C ASN gateway, PicoChip WIMAX base station and several Beceem PCMCIA and USB clients. Slice manager implemented using click router framework on ASN-gateway is responsible for maintaining the virtual network (slice) of SPs.

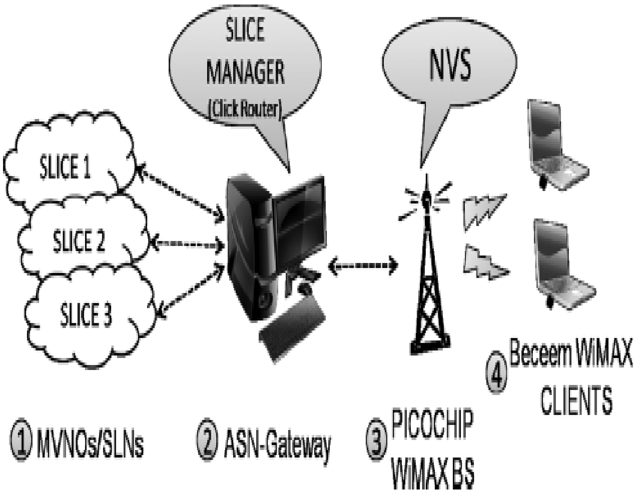

Figure 2. NVS Prototype [3]

The NVS algorithm first performs scheduling of virtual networks of SP (slice scheduling), followed by flow scheduling by SP. Slice scheduling algorithm schedules the slice with max weightage. Weight at any instant of time is defined as ratio of reserved bandwidth and exponential moving average rate (effective rate) upto that instant. If the effective rate is less than reserved resources then that slice is schedule first. The slice scheduler then allocates resources to scheduled slice based on its achieved effective rate. The achieved effective rate of slice varies depending on channel condition and scheduling policy of the slice users. These variations are accounted by considering the reference effective rate above which the slice gets its minimum reservation otherwise only fraction of reserved rate is achieved. NVS uses a concave utility function of resources allocated to the slice. This choice of utility function assumes that the marginal utility (revenue generated) of a slice decreases as its achieved resources or bandwidth increases beyond reservation. After slice scheduling, NVS allows slice to customize its flow scheduling algorithm. The NVS algorithm proves to provide isolation, customization, utilization and coexistence of slices with different requirements. This algorithm achieves virtualization of wireless resources per base station basis. But for certain deployment scenarios, virtualization may be required across a network of base stations.

Resource allocation is modeled as two stage (current and future) stochastic game which considers declared value functions (θ) by SPs and underlying channel gains (H) at each scheduling instance [4]. The value function depends on traffic state of users (g) and requested rate (r) by SP. Authors have introduced a pricing mechanism Vickrey-Clarke-Groves (VCG) using which virtual network operator (VNO) allocates a new rate (rate_alloc) and corresponding payment ($) to the user. Figure 3 gives the two stage game model which shows that all SPs provides value function θ (g, r) to pricing mechanism and VCG mechanism outputs O (rate_alloc, $).

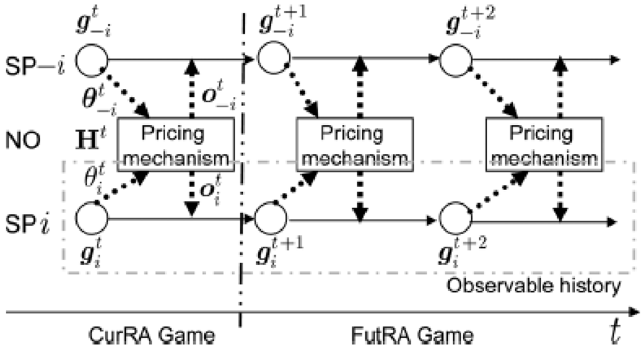

Figure 3. Two stage stochastic game [4]

The output of VCG mechanism depends on the submitted value function by SPs, and this output determines future traffic state of all SPs. SP request for a rate by considering the future congestion it will receive. But it is difficult for SP to judge the future congestion as is not aware of traffic states of other SPs and also the channel conditions. As VNO has all the information about the network, it advertises each SP's expected future congestion using this each SP can submit the value function to VNO. Thus VCG mechanism gives the optimal output for all SPs which are participating in the game at any particular instant. The result obtained shows that there is a Nash equilibrium price under various dynamic traffic and channel conditions. This paper considers convex rate region and Markovian game playing policy. But real time applications are nonconvex and playing policy has to be modeled as non-Markovian as the resource allocation is performed at finite granularity for next generation wireless system.

Long term evolution (LTE) evolved Node B (eNodeB) virtualization framework achieves the spectrum scheduling based on the service contract [5], [6]. In this framework the BSs are virtualized by using XEN hypervisor. Each virtualized BS estimate the resource requirement by calculating exponential moving average (EMA) of bandwidth over a span of time. The spectrum allocation unit in hypervisor then allocates the spectrum to virtual BS depending on EMA value, channel condition, QoS requirement and the type of contract between SP and virtual network provider (VNP). LTE eNodeB virtualization framework has considered four types of contract namely fixed guarantees, dynamic guarantees, best effort (BE) with minimum guarantees and best effort with no guarantees. Fixed guarantees contract allocates the requested fixed bandwidth to SP always independent of the usage. In dynamic guarantees contract the operator gets the bandwidth according to its actual need. The bandwidth is bounded by a maximum value that is specified by the contract. If the SP requires the maximum mentioned bandwidth, the VNP will allocate it; otherwise only the required bandwidth is allocated. The SP has to pay only for the actual used bandwidth and thus can save on the cost. Best effort (BE) with minimum guarantees contract specifies a minimum guaranteed bandwidth which will be allocated to SP at all times. As per this contract, the minimum bandwidth is always guaranteed and the additional bandwidth is assigned in a BE manner. Best effort with no guarantees type of contract allocates the bandwidth in a pure BE manner i.e. the remaining bandwidth in the network if any, is assigned to SP. Such kind of contract is suitable for non-guaranteed bit rate (nGBR) traffic. The bandwidth allocation is done by the spectrum allocation unit which first allocates the fixed bandwidth to the SPs with first contract type. Then the actual required bandwidth is allocated to SPs with second contract type which has an upper bound given in the contract. The third contract type SPs are allocated with minimum guaranteed bandwidth and the rest spectrum is distributed between the BE SPs depending on their fairness factor. As the framework allocates the bandwidth as per contract it does not achieve the optimum resource allocation.

Thus in the centralized cloud scenario, guaranteed bit rate (GBR) traffic such as voice over internet protocol (VOIP) and video is provision with demanded resources to maintain QoS while for non-GBR (nGBR) traffic such as HTTP and FTP, the resources are allocated in best effort manner. It is observed that end to end delay for GBR traffic is more in cloud network as compared to legacy network [5] [7]. For example as channel condition degrades GBR traffic requires more bandwidth which may not be available at that instant as it is already allocated to other GBR or nGBR traffic unlike legacy network which has a non-shareable fixed bandwidth. Still this end to end delay is not so large and can be tolerated.

Fluidnet is a two-step scalable framework which optimizes resource utilization to effectively handle heterogeneous users and traffic profiles [8]. This frame work employs a flexible combination of distributed antenna system (DAS) and fractional frequency reuse (FFR) depending on users and traffic profile. FFR system improves the capacity due to frequency reuse but has a small coverage. In DAS system as a single source gives the coverage over wide area, it reduces the capacity of the network. FFR and DAS systems are shown in fig. (4).

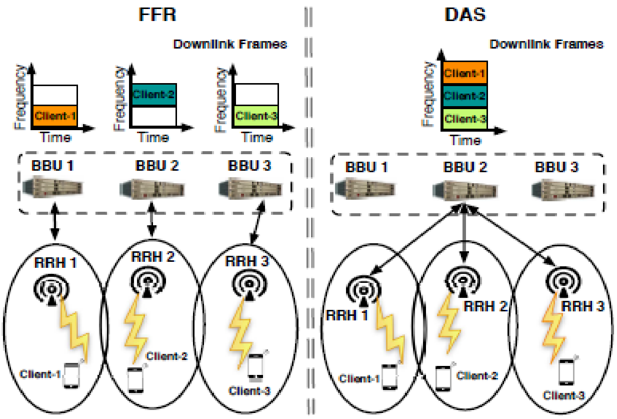

Figure 4. FFR and DAS systems [8]

For static and high traffic conditions FFR supports more traffic with desired quality of service (QoS). For mobile users and low traffic scenario DAS is implemented. Fluidnet uses hybrid configuration based on estimated resource usage from the aggregate traffic demand from each sector. The mobile traffic demand and cell-interior, cell-exterior traffic demands are used to decide the resource reservation for DAS and FFR configuration respectively. This framework adopts a two-step approach, in first step optimal DAS-FFR configuration is determined for each sector, and then further improvement in resource utilization is achieved in second step by clustering the sectors. Sectors are clustered until resource utilization of the resulting cluster cannot be improved further. Hybrid Fluidnet is compared with DAS and FFR configuration and proves to support higher traffic. It is flexible enough to adopt the variation in network conditions and also support the multiple operators and technologies. The drawback of this algorithm is that it gives priority to mobile users over fixed users.

Hybrid coordinated multipoint transmission (H-CoMP) is considered for resource allocation of delay sensitive traffic with average power and fronthaul capacity constraint in CRAN [9]. H-CoMP uses both joint processing (JP) and coordinated beamforming (CB). JP transmits the traffic jointly to user through RRHs within CoMP cluster, thus BBU transmit same traffic to different RRH using capacity constraint fronthaul. This improves spectrum efficiency at the cost of fronthaul consumption. While in CB, traffic is transmitted privately by serving RRH with interference coordination among RRHs in CoMP cluster. To achieve benefits of both schemes, the H-CoMP divides the payload of users into shared and private streams. Optimal pre-coder at RRHs and de-correlator at users are designed for both streams to maximize the mutual information for streams and to eliminate interference of other users. Authors have formulated queue aware resource allocation policy for delay sensitive traffic. The problem is formulated as constrained partially observed Markov process decision (POMDP) [10], the solution of which is obtained by equivalent Bellman equation but it has a high computational complexity. To overcome the complexity issue, linear approximation of post decision value function (function of post decision state) is used in proposed queue aware resource allocation policy with H-CoMP (QAH-CoMP). QAH-CoMP solves the optimization problem using stochastic gradient algorithm which converges to local optimum. It has been proved that the average delay per user is less in QAH-CoMP due to consideration of urgent traffic flows along with channel information for power and rate allocation.

Heterogeneous network (HetNet) which consists of small cells such as micro, pico and femto cells overlapped with the macro-cell in CRAN architecture (H-CRAN) suffers cross-tier interference. One such network is shown in fig. (5).

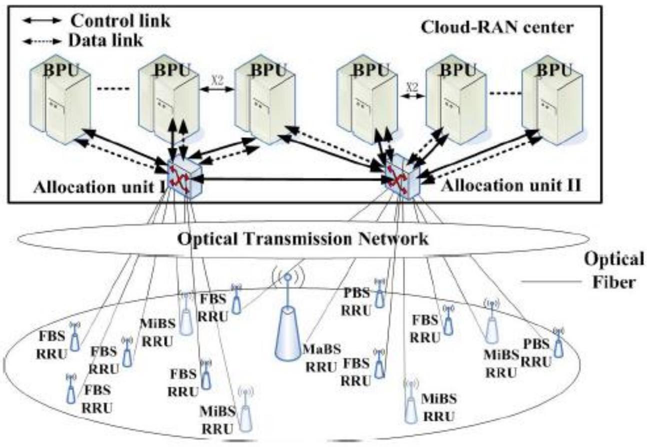

Figure 5. H-CRAN architecture [11]

To effectively deal with cross tier interference the total spectrum is shared between macro-cell and non-interfering small cells and for interfering small cells dedicated spectrum is used. A probability weighted based spectral resource allocation algorithm [11] for H-CRAN is used to find probability of small BS using shared spectrum resources. To calculate this probability cross tier interference of small cell user and macrocell user is determined. The probability will be equal to ratio of interference threshold under a certain signal to interference ratio (SIR) requirement of macrocell user to cross tier interference due to that small BS, if interference threshold of macrocell user is less than its cross tier interference, otherwise it is unity. Thus, if cross tier interference from small BS to the macrocell user increases, the probability of sharing of spectrum resources of small BS with macrocell decreases. This algorithm under centralization concept of CRAN enables more resource sharing there by increasing frequency reuse and allows severely interfering small cells to use dedicated spectrum. The algorithm outperforms clustering and power control algorithm (JFCPA) [12], the hybrid spectrum usage algorithm (HSUA) [13] and the overall fairness and efficient algorithm (OFEA) [14], in terms of frequency reuse efficiency over joint frequency bandwidth dynamic division. Also there is no outage of macrocell users as the total interference that macrocell user experienced is less than interference threshold under a certain SIR requirement of macrocell user. But this scheme gives inefficient spectrum utilization as the dedicated resources have to be assigned to interfering small cells.

H-CRAN is further extended to licenced and unlicensed access network which will be cooperating to achieve improvement in network throughput [15]. Figure 6 shows the H-CRAN in which macrocell, picocell and wi-fi access points radio resources are co-ordinately managed by centralized cooperative radio resource manager (CRRM). The interface between wi-fi and BBU is considered to be passed through a gateway which is responsible for interface matching.

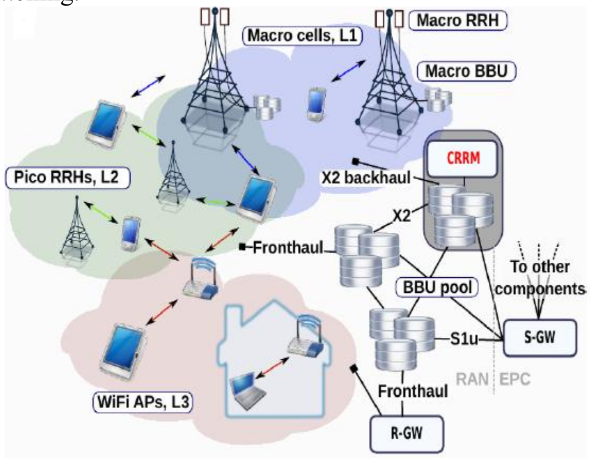

Figure 6. H-CRAN with licensed and unlicensed RAN [15]

Bifurcated uplink resource allocation scheme is used for considered scenario which splits user traffic among available multiple radio interfaces. The resources are allocated on the basis of traffic, location of user and feasible RAN connectivity. The proposed scheme achieves high throughput and fairness due to cross-radio access technology (RAT) cooperation. For the feasibility of H-CRAN with cross-RAT coordination of the network, operator needs to aggregate the wi-fi access links and cellular link. Also users have to support more than one active radio interface at a time.

Handover performance of CRAN is compared with GSM, UMTS and LTE-Advanced [16]. Inter BTS handover with the same BSC requires time advance (TA) value which is transmitted via physical information message in GSM, while for CRAN such handover is more synchronous for the BTSs in the same BBU pool. The cells belong to BTSs in a single BBU pool are synchronized and UE can directly calculate the TA by comparing the arriving time of transmitted signals from serving and target BS, thereby reducing the addition signalling overheads. Inter-NodeB handover in UMTS is soft handover which uses macro diversity with selective combining in which radio network controller (RNC) is involved. This handover in CRAN architecture becomes inter-pool softer handover which do not require RNC intervention, thus reduces signalling and increases the handover success ratio. In LTE Inter-eNB X-2 handover requires less handover prepare time in CRAN due to which too late initialization of handoff is avoided reducing the handover failure rate.

The above discussion gives the overview of different approaches used in literature for optimized resource management of CRAN network. In future more attention has to be given to elastic real time applications for the centralized CRAN along with energy efficiency which is discussed in the next section.

3. Green RAN

As mobile traffic is fluctuating throughout the day, it causes wastage of energy as all BSs are active all the time in legacy network. To reduce the energy wastage during the low traffic time, the BBU with less traffic load can be switched off and their virtual BSs (vBSs) are migrated to other BBU. Such migration becomes easy due to the centralized cloud architecture. This enhances the greener RAN concept. Thus the energy efficiency of CRAN can be improved by turning off the low traffic BBU and also by BBU and RRH power management.

To control BBU power consumption the power management unit can be used which measures the consumed power and control the power by using power management policy such as dynamic voltage and frequency scaling (DVFS) [17]. Depending on the traffic load and the BBU's processing capacity, the required number of active BBUs can be calculated, thereby achieving power saving by turning off the other BBUs. The traffic load is distributed among the active BBUs using load balancing algorithm. Equal distribution of traffic load among all active BBUs do not achieve optimum benefits as available processing capability of each of BBU is different at any particular time interval. A cooperative game-theoretic scheduling approach is used for load distribution among the BBUs, in which all BBUs negotiate for allocated workload by reporting their available processing capabilities to load balancing server (LBS) [17]. Then LBS calculate the workload shared among these BBUs under the concept of Nash Bargaining Solution (NBS). Power management unit and dynamic switching of BBUs reduces the power consumption while the NBS approach achieves load balancing between active BBUs.

In Fluidnet algorithm DAS allows one BBU to support multiple RRHs by turning off other BBUs [8] for low traffic load scenario. This achieves energy saving as compared to FFR configuration. When the low traffic BBU switch off, the migration of its traffic to other active BBU must be achieved with low service interruption time such that user should not perceive the migration. The maximum allowed interruption time for LTE handover is 300 ms. Also there should not be loss of user data during migration. Figure 7 shows the vBS migration procedure.

This vBS migration can be executed in three successive phases: migration preparation, migration execution and migration completion [18]. In migration preparation stage, BBU with enough resources to handle the migrated traffic is found first. Then the necessary system information such as cell-specific parameters is transferred from the vBS of source BBU to the vBS of target BBU. The migration execution phase switches in-phase and quadrature phase (IQ) sample path between RRH and BBU from the source vBS to the target vBS and further the necessary data is forwarded. S1 downlink path between the BBU and the core network is also switched from source vBS to the target vBS at the same time. When the IQ sample path is being switched, the source vBS starts to forward four kinds of data to the target vBS which include the following: (i) The downlink packet control and user data units buffered at the source vBS which have not been processed and delivered (ii) The downlink control and user data units that have been processed and delivered by the source vBS but have not been acknowledged by the user equipments (UEs) (iii) The data freshly arriving over the source S1 downlink path before the S1 downlink path switching is completed and (iv) The uplink receiving status which indicates the uplink packets that have been received by the source vBS and it has no opportunity to acknowledge after the IQ sample path has been switched to the target vBS. These received packets need to be acknowledged by the target vBS instead.

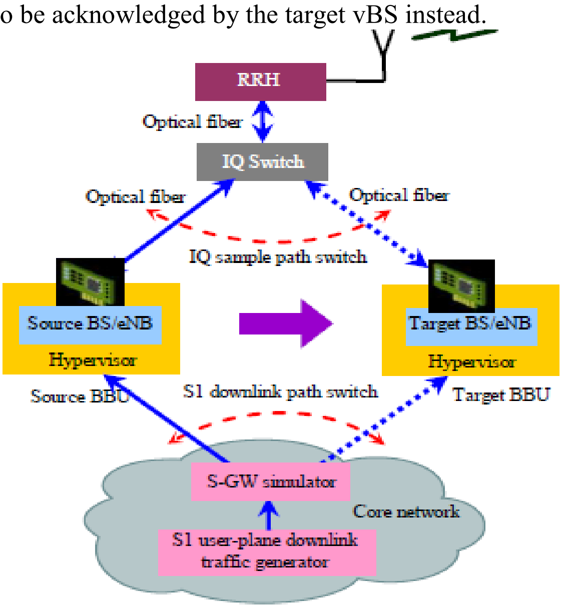

Fig. 7: vBS Migration [18]

Migration completion phase checks that all the hosted vBS on the source BBU is migrated to target BBU and after complete migration the source BBU is switch off. As vBS migration usually takes place during low traffic period, the data to be forwarded from source BBU to target BBU is less and no data loss takes place during migration process. The whole process achieves the service interruption time of tens of milliseconds for the migration of vBS to other active BBU, which needs to be reduce further as scheduling is performed on a granular scale for next generation wireless network.

BBU-RRH switching schemes namely semi-static and adaptive scheme has been proposed [19]. The semi-static scheme has a longer switching time interval in which pairs of BBU and RRH is formed to support peak hour traffic load. Adaptive scheme has a shorter switching time interval for which the resource usage at BBU is observed. In case of both the scheme, switching occurs when the BBU resource usage exceeds the upper limit or fall below the lower limit after a specified time interval. During the switching process neighbouring RRHs are preferred to be allocated to the BBU, only if it has the limit to accommodate the traffic of the RRH. As compared to static switching, adaptive scheme reduces the number of active BBU, but it has a high BBU-RRH switching rate.

The cross layer power consumption minimization is formulated as mixed integer non-linear programming (MINLP) problem [20] with constraint of BBU pool virtual machines (VMs) computational capacity, channel capacity between RRH and user, and RRHs maximum transmission power. As MINLP problem is NP-hard it is relaxed using extended sum-utility maximization (ESUM) problem which gives optimum VM computational rate and channel capacity with which ESUM problem reduces to RRH selection problem. By approximating ESUM to quasi weighted sum-rate maximization (QWSRM) problem and solving it by branch and bound (BnB) algorithm gives the global optimum VM rate and channel capacity but the complexity of BnB is high. The low complexity weighted minimum mean square error (WMMSE) algorithm is proposed which achieves local optimal rates. After achieving optimum rates, for RRH selection problem authors have proposed shaping and pruning (S-P) algorithm which finds the priorities of RRHs being chosen to be active and apply bisection search on sorted priorities to find optimal set of active RRHs. After obtaining optimal set of RRHs corresponding beamforming weights are calculated. The proposed cross layer problem has more energy efficiency but it may be unsuitable for large size network due to large channel state information (CSI) overheads and limited fronthaul capacity.

The CRAN power minimization problem through active RRH selection and the coordinated beamforming for selected RRH is addressed using group spare beamforming framework (GSBF) [21], which is a three stage approach. The network power consumption minimization problem is relaxed to obtain a weighted mixed l1/l2 norm minimization problem. The first stage of GSBF minimizes the reformulated problem to induce the group-sparsity for the beamformer. After obtaining the spare beamformer, in the second stage of GSBF the RRHs are arranged in the descending priority of switching off. The priority for RRH is decided based on channel gain, RF power amplifier gain, beamformer gain and fronthaul power consumption. After ordering RRH third step is to find the optimum number of active RRHs and the beamforming solution for the selected RRH. Bi-section GSBF algorithm uses binary procedure to obtain the active RRH, complexity of which grows logarithmically. To enhance the group sparsity for beamformer further, majorization-minimization (MM) algorithm is used. MM algorithm with conventional re-weighting also uses system prior information to improve the estimation on group sparsity. After obtaining group spare beamformer the active set of RRH is obtained by using iterative search method for the minimum power consumption which has linear complexity. Both bi-section GSBF and iterative GSBF are proved to reduce CRAN power consumption with less complexity. Iterative GSBF achieves low power consumption than bi-section GSBF, but also has the high complexity than the later. The bi-section GSBF is most suitable for large scale networks. As the GSBF is a convex relaxation of original minimization problem, the performance gap need to be analysed.

The energy efficiency (EE) maximization problem is considered for H-CRAN [22]. The soft FFR scheme is used for H-CRAN architecture in which total bandwidth is divided into two sets, one is dedicated to RRHs and another is shared between cell edge users of RRHs and users of macro BS (MBS). Shared bandwidth has the inter-tier interference constraint due to which it is used for low QoS requirement while dedicated bandwidth is used for high QoS requirements. The maximization of EE of H-CRAN is achieved through resource and power allocation subject to inter-RRH interference and inter-tier interference. For dense RRH network the sum data rate and power consumption of RRH is sufficiently larger than that of MBS. Thus H-CRAN EE optimization is approximated to EE optimization of each RRH with constraint on inter-tier interference and maximum transmission power of RRH. The formulated optimization problem is nonconvex which is solved by using iterative algorithm which converges to global optimal solution for large bandwidth. The achieved solution has high energy efficiency as compared to CRAN and one and two tier HNet, but due to approximation the EE solution does not include EE performance of the MBS.

The transmission power minimization under competitive optimality constraint and fronthaul constraint is addressed by author [23]. Competitive optimality constraint considers the fraction of informed capacity that needs to be achieved even without the information available to MS and BS, about the current fading states from the other BS at BBU. The layered transmission at MS and layered compression at BS is considered to achieve the objective without the perfect knowledge about other BS channel state information (CSI). The problem is formulated as quadratically constrained quadratic problem (QCQP) solution of which achieves local minimum. It is observed that the gain of the layered transmission decrease with increase in fronthaul capacity as it increases the quality of compressed signal.

The power optimization problem can be solved efficiently by limiting active RRHs, controlling the power of active RRHs using beamforming and controlling BBU power. The energy saving is a one of the major benefit of CRAN architecture which is feasible as multiple BBUs are in the same BBU pool and are connected through a link. The successful vBS migration is easily achieved as the capacity of the link between BBUs can handle the traffic which is comparatively low during the migration. But fronthaul link capacity may impose the constraint on the required network throughput, which is addressed in the next section.

4. Fronthaul Capacity Constraint

The high capacity link (fronthaul) is required for the efficient communication between BBU and RRH. Optical fiber, gigabit ethernet cables or wireless link can be used as fronthaul. For the wireless fronthaul millimeter wave (mm-Wave) technology can achieve data rates in Gbps [24].

In CRAN the fronthaul capacity impose a constraint on the throughput of the cellular system. The advanced technologies such as multiple input multiple output (MIMO) increases cellular throughput, but in CRAN as number of antenna at RRH increases the IQ data carried through fronthaul increases. Thus in MIMO scenario CRAN objective is maximization of sum rate of network by choosing optimal number of antennas and optimal number of users that can be served simultaneously. After pre-coding and before pre-coding methods can be used to achieve this objective of CRAN [25]. In after pre-coding approach a BBU transfers IQ-data after pre-coding data symbols with beamforming matrix, while in before pre-coding a BBU transfers data symbols without pre-coding and also beamforming weights for each data stream. Thus the required bit-rate for after pre-coding IQ-data transfer method is more as pre-coded IQ data is exchanged for each symbol between the BBU and the RRH and this bit rate depends on the number antennas used for transmission/reception at the RRH. In contrast, with before-pre-coding IQ-data transfer method, data symbols for each user are exchanged for symbol duration, but beamforming weights for each data stream are exchanged less frequently according to the channel coherence time. The optimal number of antennas for after pre-coding is directly proportional to fronthaul capacity and inversely proportional to number of bits required for IQ sample, while in before pre-coding optimal number of antennas are also inversely proportional to pre-coding matrix update frequency. As in before pre-coding the information for pre-coding is less frequently transmitted, it is more efficient for low speed mobile scenario. The required fronthaul rate for low speed users with after pre-coding is much higher than the highest link rate option (9.82 Gbps) in the current common public radio interface (CPRI) specification [26]. For high speed users before pre-coding gives the worse result as now beamforming weights needs to be updated frequently. Thus by choosing optimum number of antennas and number of simultaneous users, the network capacity is maximized by using appropriate pre-coding method depending on mobile system environment under the constraint of fronthaul capacity. By IQ compression method the sum rate of the system can be improved further.

The compression of signals on fronthaul is performed using Wyner-Ziv coding and single user coding at RRH to achieve an optimal sum capacity for uplink [27]. The Wyner-Zive scheme considers the statistical correlation of the received signals at different RRHs for compression process, while in single user coding each RRH simply quantizes its received signals using a vector quantizer. The problem is formulated for the optimum quantization level to achieve high weighted sum rate with fronthaul capacity constraints. The alternative convex optimization (ACO) approach is used which converges to a local stationary point of the weighted sum rate maximization problem. Although ACO achieves the locally optimal quantization noise level, its implementation in fast fading environment and multiplexed system is computationally intensive. Thus to obtain the optimal quantization noise level with less computational complexity, the authors have proved that in the high signal to quantization noise ratio (SQNR) regime, the quantization noise level can be set proportional to the background noise level. This is approximately optimal for maximizing overall sum rate for both Wyner-Ziv coding and single user coding. Authors have also proved that, setting the quantization noise level proportional to the background noise also achieves optimal results for H-CRAN. It is proved that the CRAN with Wyner-Ziv coding and single user coding compression on fronthaul achieves the higher uplink user rates as compared to the conventional cellular system. The comparison between Wyner-Ziv coding and single user coding scheme shows that the Wyner-Ziv is superior, however as the sum fronthaul capacity becomes larger, the gain due to Wyner-Ziv coding diminishes.

H-CoMP transmission mode is used for different fronthaul capacity in CRAN to maximize average net throughput of the network under the fronthaul constraints [28]. To implement H-CoMP the semi-dynamic clustering for user is performed by BBU based on large scale channel information of different BS reported by user. The overheads of clustering are less in case of large scale fading as compared to small scale fading, thus improves average net downlink throughput. Cluster of users contains master BSs which serve user (JT mode) and co-ordinated BSs which avoid inter-channel interference (ICI) for user (CB mode). The number of users serves and co-ordinated by BS is restricted by number of antennas at BS. Once the clusters are formed each BS's pre-coder calculates weights based on downlink channel information. The zero-forcing pre-coder is used to cancel ICI for coordinated users of BS and ensures co-phase signals coming from master BSs to user for constructive combination. To reduce the large scale channel reporting overheads the user choose the BS which has a high larger scale channel gain while neglecting the BS which are far away from user. This neglected interference effect on throughput estimation is removed by using weak interference estimation mechanism. After calculating pre-coding weights, BBU shares data of user to different BSs in its cluster. The proposed semi-dynamic hybrid CoMP achieves same average net throughput as optimal H-CoMP which uses exhaustive search for clustering which increases its complexity.

The weighted sum rate (WSR) maximization for user centric clustering strategy under per BS fronthaul constraint is addressed [29] using reweighted l1-norm. The user centric clustering has two approaches namely dynamic and static clustering. Dynamic clustering changes the cluster of serving BSs of user over a time-frequency slot while the static clustering keeps BSs cluster of a user fixed over a time which only changes with change in location of user. The dynamic clustering uses the fronthaul resources more efficiently than the static but it has a more signalling overheads. The WSR maximization problem is solved by formulating BS fronthaul constraint by weighted l0-norm which is the reformulated to l1-norm. The reformulated problem is equivalently expressed as weighted minimum mean square error (WMMSE) problem which is convex and solved efficiently using block coordinated decent method. To reduce the complexity of the algorithm the iterative link removal and iterative user pool shrinking techniques are proposed for dynamic clustering approach. In iterative link removal technique the BSs with the transmit power below threshold are removed from cluster. The iterative user pool shrinking technique reduces the number of users which achieves the negligible rates. The dynamic clustering determines the cluster size for each user in each slot and then determines jointly the beamforming vector and user scheduling. Unlike dynamic approach, static approach selects the serving users at a scheduling interval for the fixed cluster. To form static cluster for user the two heuristic load balancing schemes are introduced namely maximum loading based static clustering and biased signal strength based static clustering. First method selects cluster of BSs for user depending on the threshold on received signal strength difference while later considers biased received signal strength difference for each user. Dynamic and static clustering schemes improves the performance of low rate users (cell edge users) as compared to baseline disjoint clustering scheme [30] which gives the performance gain for high rate users.

Limited work is published in literature on the fronthaul capacity issue which is an open research area for the future work.

5. Existing Technologies and Future Research Direction

The existing commercial products support the different levels of virtualization. VANU MultiRAN [VAN] [31] is a hardware level virtualization which supports multiple virtual base stations all running on a single hardware platform. The multiple virtual base stations can share the antennas, hardware platform and the backhaul. But this solution do not support spectrum sharing among different operator slices, thus only helps to reduce cost and energy of multiple BSs of different operator by supporting them on single hardware.

China mobile research institute, which is the inventor of CRAN, performed field trials to evaluate the performance of CRAN in terms of centralization and fronthaul [32]. CRAN trial was carried for 2G and 3G in which it was observed that power reduction of 41% due to shared air-conditioning. With the mentioned benefits of centralization the network performance has improved with respect to call drop rate and throughput. For fronthaul, CPRI compression with ratio 2:1 and single fiber bi-directional (SFBD) technology was used which proves the fourfold saving of fiber resource as compared to dark fiber solution. The performance degradation due to CPRI compression with respect to handover, throughput, data and control plane latency is also negligible. It has proved that the wavelength-division multiplexing (WDM) also achieves less fiber consumption than that of dark fiber with small processing delay and the quick protection switching capability. Field trial also proved the 20-50% throughput gain through uplink CoMP in CRAN.

LightRadio [33] by alcatel-lucent is a portable radio which supports multiple cellular standards such as wideband code division multiple access (W-CDMA) and LTE technologies and their evolution. The SP can start with a W-CDMA and gradually can upgrade to LTE standards by performing software reconfiguration. LightRadio supports both, in radio (RRH) baseband processing and centralized pooled baseband processing. Based on the baseband processing it supports latency insensitive, low data rate fronthaul and latency sensitive, high data rate fronthaul. This centralized multi-band radio network achieves cooperative performance gain and also energy and cost saving. LightRadio network increases the capacity and can also be used as hotspots to support massive capacity.

The alcatel-lucent with china mobile has come up with virtualized radio access network (vRAN) [34] which is a virtualized, centralized cellular network with network functions virtualization (NFV) technology. Currently NFV is used for a core network and IP-based Multimedia Service (IMS) functions. The trial of vRAN was carried out at Beijing's Tisinghua University demonstrated the flexibility, scalability, cost and energy efficiency of the vRAN.

Antenna integrated radio unit (AIR) [35] by Ericsson is also a multi-standard solution which supports both 3G and 4G. Field trials proved a reduction of integration and installation time of up to 30% and the reduction in power consumption of up to 42% [35].

The use of the existing commercial solutions are still restricted to single SP supporting multiple standards and processing those standards in centralized pool. These solutions can be used over the multiple SPs to further enhance the benefits.

CRAN architecture will lead to development and implementation of new generation systems much cheaper, faster and more flexible, with centralized scheduling, coordination and network maintenance. But the benefits of CRAN will be successfully achieved only after dealing with the new challenges of this centralized architecture. Major requirements of this architecture are real time processing of vBS pool, efficient cooperative algorithms for resource and energy optimization and high bandwidth low latency fronthaul. Virtual BS pool which is implemented on general purpose processor (GPP) has low performance in terms of real time processing and energy efficiency. Also optical fibre as a fronthaul does not achieve cost benefit. Complexity of resource optimization algorithms is also high due to centralized and cooperative approach. Thus improvement in performance of GPP, efficient fronthaul both in terms of data rate and cost and optimal algorithms to manage network resources are the key steps in deployment of CRAN concept [36]. To achieve the maximum benefits from CRAN architecture the research can be carried out in the following directions:

1. **Development of radio resource management (RRM) algorithm**: Tunable RRM algorithm needs to be design which should have the flexibility to adapt to the degree of centralization, based on the density and load distribution of small cells taking into account also other parameters like energy efficiency and other backhaul/access network related parameters. For highly dense and populated network more centralized solutions are prefer to cope up with the increasing signaling overhead that could be imposed by the high density of small cells. Thus RRM algorithm must be adaptive to all traffic loads of various types.
2. **Efficient Fronthaul Solution:** The fronthaul capacity and cost constraint need to be considered while designing CRAN. Implementation of heterogeneous fronthaul with high capacity links (optical fiber) and low capacity links (wireless links) has to consider [37]. Low capacity link will reduce cost but has poor performance of latency and throughput as compared to high capacity links. Optimization w. r.t. capacity and cost can achieve by using heterogeneous fronthaul. The wireless fronthaul using mm-Wave can achieve the high capacity but these signals suffer high attenuation. The high degradation of mm-Wave needs to be overcome using efficient antenna design and using MIMO approach.

   The interface standardization will have to be developed for the heterogeneous fronthaul. Transmitter and receiver cooperation algorithm need to be flexible enough to handle different fronthaul characteristics with respect to throughput, latency, and reliability. The data compression schemes can be explored further to handle the capacity constraints of fronthaul. Also some functionality within the BBU can be distributed among the local RRH to reduce the traffic on fronthaul.
3. **Self-organized network**: For given dynamic service requirements, user mobility, fronthaul constraints and backhaul network conditions, access and core networks need to be jointly optimized and managed in an autonomic way to reduce the operational expenditures (OPEX), while enabling fast response and energy efficiency. Such self-organized network needs to be designed for intelligent load balancing, and the admission and congestion control.
4. **Self-constructing network**: In conventional network each RRH transmits cell specific information, synchronization signals for terminal. In dense small cell scenario number of RRHs transmit the same information as they belong to same geographical location. This increases interference, overloads the network and causes power wastage. In centralized CRAN architecture broadcast and synchronization information can be schedule among the RRHs belong to same geographical area [38], thereby reducing transmission of replicated information. Thus smart self-constructing network will help to reduce unnecessary signaling in the network.
5. **Energy aware RAN**: Energy efficient algorithm and lossless data migration techniques during sleep mode of BBU need to be explored further and implemented.

The CRAN framework is still in nutshell and there are many open research directions need to be explored to optimize the benefits of the framework.

6. Conclusion

CRAN is a promising technology which is a major evolution over existing confined traditional architecture. Centralized CRAN architecture allows dynamic spectrum sharing and coordination and cooperation among the different SPs, achieving many benefits and fulfils the need of future network. Along with the benefits, centralization also increases the complexity of network optimization. This paper presented a survey of different resource management and energy efficient approach. The paper also surveyed network throughput enhancement techniques under fronthaul capacity constraints.

The intelligent and self-reconfigurable CRAN architecture need to develop which opens many future challenges.

References

[1] CISCO, VNI Mobile Forecast Highlights, 2012 – 2017. [Online] Available: http://www.cisco.com/web/solutions/sp/vni/vni_mobile_forecast_highlight/index.html

[2] China mobile research institute, "C-RAN: The road towards green RAN", in *C-RAN International Workshop*, Beijing, China, Apr. 23rd 2010.

[3] R. Kokku, R. Mahindra and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks", *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, Oct. 2012.

[4] F. Fu and U. Kozat, "Stochastic game for wireless network virtualization", *IEEE/ACM Transactions on Networking*, 2012.

[5] Y. Zaki, "Future mobile communications: LTE optimization and mobile network virtualization," Ph.D. dissertation, Dept. Physics and Electrical Engineering, Bremen Univ., Iraq, 2012.

[6] X. Wang, P. Krishnamurthy, and D. Tipper, "Wireless network virtualization", *International Conference on Computing, Networking and Communications*,2013.

[7] M. Li, L. Zhao, and et al., "Investigation of network virtualization and load balancing techniques in LTE networks", *IEEE 75thVehicular Technology Conference (VTC Spring),* 2012.

[8] K. Sundaresan, M. Arslan, S. Singh, S. Rangarajan and Srikanth V., "FluidNet: A flexible cloud-based radio access network for small cells", *MobiCom '13 Proceedings of the 19th annual international conference on Mobile computing & networking*, pp. 99-110, 2013.

[9] J. Li, M. Peng, A. Cheng,Y. Yu, and C. Wang, "Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks", *IEEE Systems Journal*, Issue: 99, Nov. 2014.

[10] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. II, Massachusetts: Athena Scientific, 2007.

[11] Z. Wang, H. Li, and et al., "Probability weighted based spectral resources allocation algorithm in Hetnet under Cloud-RAN architecture", *1stIEEE ICCC International workshop on internet of things*, 2013.

[12] H. Li, X. Xu, D. Hu, X. Qu, X. Tao, and P. Zhang, "Graph method based clustering strategy for femtocell interference management and spectrum efficiency improvement," *Journal of Communications and Networks*, Dec. 2011, vol. 13, no. 6, pp. 664-677.

[13] Y. Bai, J. Zhou and L. Chen, "Hybrid spectrum usage for overlaying LTE macrocell and femtocell," *Proc. IEEE Global Telecommunications Conference*, Dec. 2009, pp. 1-6.

[14] J. Lu, W. Zheng, T. Su, and X. Wen, "Interference mitigation spectrum allocation for energy efficient OFDMA femtocell networks," *Proc. International Conference on ICCSNT,* Dec. 2012, pp. 252-256.

[15] M. Gerasimenko, D. Moltchanov, and et al., "Cooperative radio resource management in heterogeneous cloud radio access networks", *IEEE Access,* Vol.3, Apr. 2015.

[16] L. Liu, F. Yang, R. Wang, Z. Shi, A. Stidwell, and D. Gu, "Analysis of handover performance improvement in cloud-ran architecture", *7th International ICST Conference on Communications and Networking in China (CHINACOM)*, Aug. 2012.

[17] Z. Kong, J. Gong, C. Xu, K. Wang, and J. Rao, "eBase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network", *IEEE International Conference on Communications,* 2013.

[18] C. Wang, Y. Wang, G. Chaohua, W. Yan, C. Liyu, and L. Qinglin, "A study on virtual BS live migration – a seamless and lossless mechanism for virtual BS migration", *IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications: Mobile and Wireless Networks*, 2013.

[19] S. Namba, T. Warabino and S. Kaneko, "BBU-RRH switching schemes for centralized RAN", *7th International ICST Conference on Communications and Networking in China (CHINACOM),* Aug. 2012.

[20] J. Tang, W. Tay and T. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network", *IEEE Transactions on Wireless Communications*, May 2015.

[21] Y. Shi, J. Zhang and K. Letaief, "Group Sparse Beamforming for Green Cloud-RAN", *IEEE Transactions on Wireless Communications*, Vol. 13, No. 5, May 2014.

[22] M. Peng, K. Zhang,J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks", *IEEE Transactions on Vehicular Technology*, Dec. 2014.

[23] S. Park, O. Simeone, O. Sahinand S. Shamai, "Robust Layered Transmission and Compression for Distributed Uplink Reception in Cloud Radio Access Networks", IEEE Transactions on Vehicular Technology, Vol. 63, No. 1, January 2014.

[24] K. Huang, Z. Wang, *Millimeterwave communication systems,* Wiley-IEEE Press, 2011.

[25] S. Park, C. Chae, and S. Bahk, "Before/After pre-coding massive MIMO systems for cloud radio access networks", *JOURNAL OF COMMUNICATIONS AND NETWORKS*, Vol. 15, No. 4, Aug. 2013.

[26] CPRI specification v5.0 (2011-09-21), [Online]. Available: http://www.cpri.info /downloads/CPRI_v_5_0_2011-09-21.pdf

[27] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network", *IEEE Journal on Selected Areas in Communications,* Volume: 32, Issue: 6, 2014, pp. 1295 – 1307, 2014.

[28] D. Liu, S. Han, C. Yang, and Q. Zhang, "Semi-dynamic user-specific clustering for downlink cloud radio access network", *IEEE Transactions on Vehicular Technology*, May 2015.

[29] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network", *IEEE Access Special Section on Recent Advances in Cloud Radio Access Network,* Vol. 2, Nov. 2014.

[30] S. Kaviani, O. Simeone, W. Krzymien and S. Shamai, "Linear pre-coding and equalization for network MIMO with partial cooperation", *IEEE Transactions on Vehicular Technology*, Vol. 61, No. 5, June 2012.

[31] "Vanu networks", Vanu, Inc., Cambridge, MA, 2011 [Online]. Available: http://www.vanu.com/

[32] C. Lin, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent Progress on C-RAN Centralization and Cloudification", *IEEE Access Special Section on Recent Advances in Cloud Radio Access Network,* Vol. 2, Sept. 2014.

[33] lightRadio™ Technology Overview, [Online]. Available: https://techzine.alcatel-lucent.com/lightradio^{TM}-technology-overview

[34] How vRAN is helping future-proof mobile networks, [Online]. Available: https://techzine.alcatel-lucent.com/how-vran-helping-future-proof-mobile-networks

[35] Ericsson unveils antenna integrated radio unit: AIR, Ericsson Press Release, Feb. 2011.

[36] M. Hadzialic, B. Dosenovic, M. Dzaferagic, and J. Musovic, “Cloud-RAN: innovative radio access network architecture”, *ELMAR, 55th International Symposium*, 2013.

[37] D. Sabella, P. Rost and Y. Sheng, “RAN as a service: Challenges of designing a flexible RAN architecture in a cloud based heterogeneous mobile network”, *Future Network & MobileSummit 2013 Conference Proceedings,* 2013.

[38] M. Webb, Z. Li, P. Bucknell, T. Moulsley, and S. Vadgama, “Future evolution in wireless network architectures: towards a cloud of antennas”, *IEEE Vehicular Technology Conference (VTC Fall),* 2012.