User QoS-aware Adaptive RF Chain Switching for Power Efficient Cooperative Base Stations

Ranjini Guruprasad and Sujit Dey, *Fellow, IEEE*

***Abstract*—We propose an user Quality of Service (QoS) and base station (BS) resource utilization aware radio frequency (RF) chain switching technique among cooperating BSs, termed cooperative RFSnooze (Co-RFSnooze), to improve the power efficiency of cellular networks. The key idea is to maximize the number of RF chains that can be switched off in a cluster of neighboring BSs that have overlapping coverage areas. To achieve this, we propose to jointly explore the individual BS resource space consisting of number of RF chains, frequency blocks and time slots and the user association (UA) space formed by users located in coverage areas of multiple BSs in the cluster. Specifically, we formulate the problem to minimize the sum of average power consumption of cluster of BSs in a transmission frame with users' QoS and BS resource utilization as constraints to be satisfied. We then propose a heuristic iterative algorithm to solve the optimization problem. Simulation results based on real dataset demonstrate that the proposed Co-RFSnooze technique can achieve up to 44% savings in average cluster power consumption in a transmission frame while satisfying the users' QoS and BS utilization constraints.**

***Keywords*-User QoS, BS resource adaptation, BS cluster, User association adaptation, Adaptive RF chain switching, Cluster power consumption.**

I. Introduction

By 2022, the expected number of mobile subscriptions and the resulting mobile traffic is expected to reach 8.9 billion subscriptions and 69 Ebytes respectively [2]. To cater to the explosive growth in mobile data subscriptions and traffic, it is estimated that the total number of base stations (BSs) in cellular networks all over the world will grow to 11.2 million by 2020 [3], a 47% increase compared to the number of BSs deployed in 2014. Further, deployment of massive number of antennas at BSs is seen as a promising paradigm to increase data rates [4]. This is expected to increase the electricity consumption and thereby, decrease the energy efficiency of cellular networks [4]. Specifically, the electricity consumption of BSs which constitutes 80% of electricity consumption of cellular networks is estimated to increase from 84TWh to 109TWh (38% increase) if measures are not taken to reduce the power consumption of BSs. The increasing electricity consumption has two effects - (a) the carbon equivalent emissions is estimated to increase to 235 Mto CO_{2e} by 2020 (a 37% increase from 2014) [3] and (b) the electricity bill which currently contributes to 10-15% of the operating expenses in developed markets and about 50% [5] in developing markets will further increase. Hence, increasing the power efficiency of base stations becomes a critical requirement to reduce growing operating cost for mobile operators and to comply with the trending global desire to reduce energy consumption and carbon footprint, and increase sustainability.

The authors are with the Electrical and Computer Engineering Department, University of California San Diego, CA 92093 USA. e-mail: (rgurupra@ucsd.edu, dey@ucsd.edu).

Part of this work has been published in the Proceedings of IEEE International Conference on Communications (ICC), London, June 2015 [1].

Amongst many components of the BS, the power amplifier (PA) in RF chain consumes about 65% [6] of the total power consumption in the BS. Further, multi-input multi-output (MIMO) BS providing high data rates and enhanced coverage uses multiple RF chains which increase the contribution of RF chain power consumption. Consequently, to reduce BS power consumption, it is vital to develop techniques that can lower RF chain power consumption.

The total power consumption due to RF chains is determined by the number of active RF chains, transmission power, transmission bandwidth and duration of transmission required to satisfy the Quality of Service (QoS) i.e., throughput and block error rate (BLER) requirements of the users. Given the user association (UA), there may exist multiple combinations of the above-mentioned BS resources that satisfy the users' QoS requirements and which result in varying levels of BS resource utilization and RF chain power consumption [1].

Moving from single BS to cluster of BSs which have overlapping coverage areas, there may be multiple users located in the coverage area of more than one BS. This implies that there may exist multiple combinations of UA across the cluster BSs which will satisfy the QoS requirements of all the users associated with the cluster BSs. Different combinations of UA can result in different BS resource utilizations and hence RF chain power consumption.

In this paper, we propose a cooperative adaptive RF chain switching technique which explores the BS resource and UA spaces to maximize the number of RF chains that can be switched off to minimize RF chain power consumption and thereby power consumption of the BSs in the cluster. While trying to adapt the BS resources and UA, the proposed technique ensures that individual BS utilization constraints are not violated and QoS requirements of all the users in the cluster are satisfied.

A. Related Work

In this section, we will briefly describe prior work related to BS resource and UA adaptation to achieve adaptive RF chain switching (RFS) and power efficient operation of cellular networks. The relevant techniques are grouped in to three categories based on (a) the number of BSs considered for applying the BS on/off, BS resource and UA adaptation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

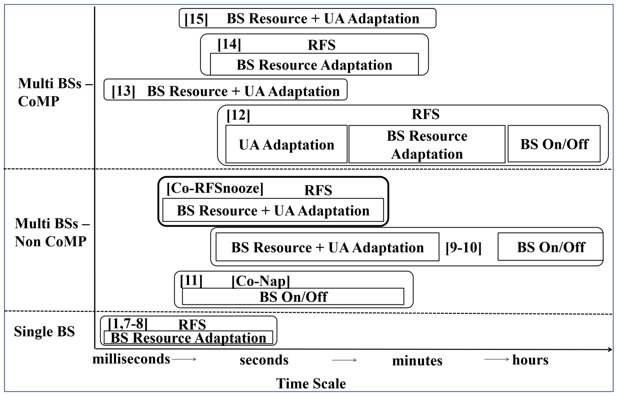

Fig. 1. Comparison of related work with the proposed Co-RFSnooze technique

techniques and (b) the use of coordinated multi-point (CoMP) transmission. Note that, though BS on/off switches RF chains, it is not adaptive as BS on/off either switches on or off all RF chains. Further, in each category, techniques are distinguished based on time scale of operation. We will refer to time scales of milliseconds to minutes as short time scale and tens of minutes to hours as long time scale. The above described grouping is shown in Fig. 1.

We will first discuss the techniques applicable to a single BS as shown in the bottom row of Fig. 1. The technique (termed Min-Cost in [1] and RFSnooze in this paper) proposed in the preliminary version of this paper [1] adapts the number of RF chains, time slots and frequency blocks while satisfying both the users' throughput and BLER requirements as well as BS utilization constraints. Authors in [7] propose data rate, power, RF chain and subcarrier allocation in a manner that maximizes the energy efficiency of data transmission of a single BS. The technique proposed in [8] jointly maximizes transmitter and receiver energy efficiency of a single BS and associated users. In contrast to the above single BS techniques, the proposed short time scale Co-RFSnooze technique is applicable to cluster of cooperating BSs. It extends [1] to jointly adapt the individual BS resources as well as the UA of all the cluster users (Section IIID) to maximize the number of RF chains that can be switched off in the entire cluster and minimize the cluster power consumption. We will next discuss the techniques which are applicable to a cluster of cooperating BSs that do not use CoMP transmission (middle row, Fig. 1).

Dynamic BS on (active)/off (inactive) techniques switch BSs on or off based on number of associated users [9] and the estimated savings in power consumption due to switching off of BSs [10]. The above techniques switch off all the components of a BS which takes tens of minutes and can be classified as a long time scale operation. Though short time scale operations of BS resource and UA adaptation are applied to the subset of active BSs, long time scale switching off of BSs could potentially lead to coverage holes. Coverage holes are a major concern for the operators as a user in the coverage hole will not receive coverage. In contrast, our proposed approach adapts BS resources and UA on a short time scale enabling finer tracking of the BS load and finer control on BS power consumption without degrading coverage capabilities.

The Co-Nap technique proposed in [11] implements short time scale BS on/off by adapting the number of "nap" (sleep) time slots for the cluster BSs in a coordinated manner. As all the BS RF chains are switched off in the "nap" time slots, it reduces BS power consumption. Unlike the Co-Nap strategy which adapts only the on/off pattern of BSs, the proposed Co-RFSnooze technique jointly adapts BS resources and UA to achieve adaptive RFS. We will demonstrate in Section IVB that this joint adaptation achieves higher power efficiency compared to Co-Nap.

Next, we will discuss techniques that are applicable to cluster of cooperating BSs using CoMP transmission (top row, Fig. 1). The long time scale technique in [12] determines the BS and RF chain on/off pattern, UA and power allocation and the short time scale technique in [13] exploits the varying delay tolerance of users to enable time slot based BS sleep. The throughput requirements of the users associated with the inactive BS in [12]-[13] are met through CoMP transmission by the active BSs in the cluster. The authors in [14] propose a resource allocation algorithm for full-duplex, distributed antenna, multi-user communication network that minimizes the power consumption of cluster of BSs by dynamically switching off RF chains while satisfying the QoS requirements of downlink and uplink users. The above techniques require sharing of the channel state information (CSI) and data of all the users in the cluster via the backhaul to compute the multi-cell precoding matrix to perform CoMP transmission. The proposed Co-RFSnooze technique does not utilize CoMP transmission and instead proposes novel heuristics and combination of centralized-decentralized framework that requires sharing of only the user QoS and association information to significantly reduce the communication via the backhaul. As shown in Fig. 5b (Section IVB), there are 270 users in the cluster during high load and the techniques [12]-[14] will require sharing CSI information and data of all the 270 users whereas the proposed technique requires user QoS and association information of only 35 users (users transferred shown in Fig. 6b).

The technique proposed in [15] determines the BS-user association for CoMP transmission and performs joint spectrum and power allocation to minimize the total cluster transmission power. However, [15] does not dynamically switch off RF chains and always maintains them in the on state. In contrast, the proposed Co-RFSnooze technique performs BS resource and UA adaptation to dynamically switch off RF chains in the cluster. This can potentially result in higher power savings compared to [15] which always switches on all the RF chains (demonstrated in Section IVB by significant savings compared to All-On/Co-Nap which switches on all RF chains).

From the above description of the prior art, to the best of our knowledge, this is the first work

- that dynamically switches RF chains in a cluster of cooperating BSs by jointly adapting BS resources and UA on a short time scale to minimize the average cluster power consumption in a transmission frame.
- that jointly adapts BS resources and cluster UA in a

TABLE I
SUMMARY OF NOTATIONS USED

$\mathcal{B}, BW$	Set of BSs in the network, Transmission bandwidth of BS $b \in \mathcal{B}$
S, R	Maximum number of RF chains at BS and user
P_b^{Tx}, P^{Max}	Transmit power and maximum transmit power of BS b
t^F	Duration of frame
T, T^A, T^I	Number of time slots in a frame, Number of active and idle time slots in a frame
t^O, t^{Sw}	Duration over which all RF chains are off in a frame, RF chain switching duration in a frame
S_t^A, S_t^O, S^{Sw}	Number of active and off RF chains in time slot t, Number of RF chains switching state in a frame
J, ψ_{st}	Number of frequency blocks in time slot $t \in T$, Frequency utilization of RF chain s in time slot t
m, M	Transmission mode and set of all transmission modes
$s_{ib}(m)$	Number of BS RF chains allocated by BS b to the i^{th} user for mode m
$r_{ib}(m)$	Number of RF chains allocated by i^{th} user associated with BS b for mode m
$d_{ib}(m)$	Number of independent data streams received by i^{th} user associated with BS b for mode m
$\gamma_i, BLER_i^{Th}$	Throughput requirement of i^{th} user, Upper bound on BLER requirement of i^{th} user
$\boldsymbol{H}_{ib}, SINR_{ib}$	Channel matrix between i^{th} user and BS b, Signal to interference noise received by i^{th} user from BS b
$TP_{ib}, BLER_{ib}$	Throughput provided by BS b to i^{th} user, BLER provided by BS b to i^{th} user
I_b	Set of users associated with BS b
I_b^{NT}, I_b^T	Set of non-transferable and transferable users associated with BS b
$I_b^{T\sim}$	Subset of I_b^T users associated with BS b that require the same set of RF chains and time slots as users I_b^{NT}
P^I, P^O, P^{Sw}	Idle and off power consumption of BS, PA switching power
Δ_p	Power gradient
P_b, P_C	Average power consumption of BS b in a frame, Average cluster power consumption in a frame
$C, \|C\|$	Set of cluster BSs and number of cluster BSs in cluster C
I_C, I_C^{NT}, I_C^T	Set of users in cluster C, Set of non-transferable and transferable users in cluster C
BSU, k_{bi}	BS-user matrix of size $\|C\|x\|I_C\|$, entry in BSU matrix of BS b for i^{th} user
E_i	Set of BSs that satisfy i^{th} user's mode SINR threshold
g, E	Transferor BS, set of transferee BSs
RFU	Number of active RF chains to users ratio

manner that the cluster user's QoS requirements and the BS resource utilization constraints are satisfied.
- that does not require BS switching and expensive CoMP data transfer and matrix computations to adaptively switch RF chains in a cluster of cooperating BSs.

The rest of the paper is organized as follows. Table I summarizes the notations used. Section II describes the system model and the optimization problem. In Section III, we propose a heuristic algorithm to solve the optimization problem. In Section IV, we provide simulation results under a practical configuration. Finally, we conclude the paper in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Network, Channel and User QoS Models

Consider the downlink communication in MIMO-Orthogonal Frequency Division Multiple Access (OFDMA) cellular network with set of BSs $\mathcal{B}$ as shown in Fig. 2. The overall bandwidth BW is divided in to J equally sized frequency blocks and the transmission frame of duration t^F is divided in to T equally spaced time slots, each of duration $\frac{t^F}{T}$. The maximum number of RF chains that can be active at BS $b \in \mathcal{B}$ and each user device are S and R respectively. We will define a transmission mode m as $m \triangleq (s(m), r(m), d(m))$ where $s(m) \in [1, S]$ is the number of BS RF chains required for mode m, $r(m) \in [1, R]$ is the number of RF chains required at the user device and $d(m) = min(s(m), r(m))$ is the number of independent data streams transmitted by mode m. We assume single-input single-output (SISO) and Single User-MIMO (SU-MIMO) including spatial multiplexing (SM) and spatial diversity (SD) modes for transmission. We will denote the set of all possible transmission modes as M. In this paper, mode selection is done once every transmission frame and the mode $m_{ib} \in M$ selected for the i^{th} user by BS b does not change within time slots of a frame. Hence, the number of RF chains $s_{ib}(m)$ allocated by BS b to the i^{th} user, number of RF chains $r_{ib}(m)$ allocated by the i^{th} user device and the number of independent data streams $d_{ib}(m)$ received by the i^{th} user remains identical for all the active time slots of the frame. Let I_b denote the set of users associated with BS b and $I_b^T \subseteq I_b$ denote the subset of 'transferable' users who are in coverage area of BSs $b^\sim \in \mathcal{B} \setminus b$ in addition to being in the coverage area of BS b. For cooperative RF chain switching, we propose to adapt the UA of such transferable users which lie in the coverage areas of multiple BSs. This motivates us to consider group or cluster of BSs $C \subseteq \mathcal{B}$ having overlapping areas of coverage enabling cooperation and UA adaptation. In this paper, we adopt the network centric clustering of BSs wherein BSs are grouped together statically based on network planning considerations [16]. Like used extensively in related research [11] and [17], we assume that the set $\mathcal{B}$ can be divided in to disjoint clusters of BSs and the size of each cluster is $|C|$ where $|X|$ denotes the cardinality of set X. We also assume that all the BSs in the cooperative cluster can communicate with each other via the X2 interface. We assume block fading channel between BS b and the i^{th} user over the entire bandwidth (J frequency blocks) in a frame (T time slots) represented by the complex channel matrix $\boldsymbol{H_{ib}} \in C^{r_{ib}xs_{ib}}$ of rank $A \leq d_{ib}$. The noise at each user's receiver is assumed to be additive white Gaussian with zero mean and variance σ^2. We assume that the user's channel state information (CSI) including channel quality information (CQI) and Rank Indicator (RI) is available at the BS.

Assuming that the transmit power P_b^{Tx} of BS b is equally divided over all frequency blocks and transmit antennas, the signal to interference-noise ratio (SINR) received by the i^{th} user is

$$SINR_{ib} = \frac{P_b^{Tx}}{Js_{ib}} \cdot \frac{\mathbf{H}_{ib}\mathbf{H}_{ib}^H}{\sum_{b^\sim \in \mathcal{B} \setminus b} P_{b^\sim}^{Tx} \mathbf{H}_{ib^\sim}\mathbf{H}_{ib^\sim}^H + \sigma^2} \quad (1)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

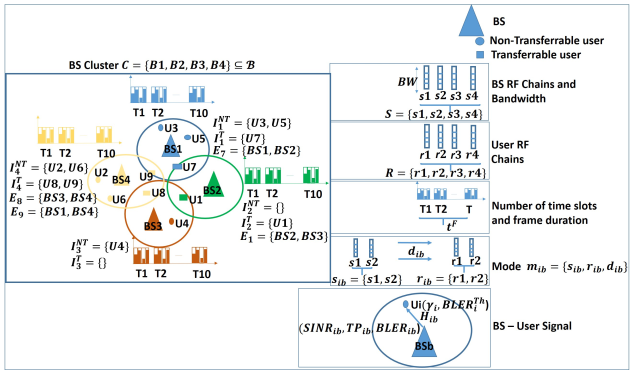

Fig. 2. System Block Diagram

The throughput TP_{ib} from BS b to i^{th} user is given by

$$TP_{ib} = \frac{BW}{JT} \sum_{t=1}^{T_{ib}} J_{tib} \log_2[\det\{I_{r_{ib}} + SINR_{ib}\}] \tag{2}$$

where T_{ib} is the number of time slots and J_{tib} is the number of frequency blocks assigned in time slot $t \in [1, T_{ib}]$ by BS b to the i^{th} user and $I_{r_{ib}}$ is a $r_{ib} x r_{ib}$ identity matrix. The $BLER_{ib}$ achieved for the i^{th} user depends on the BS transmit power P_b^{Tx}, channel $\boldsymbol{H_{ib}}$, and the mode m_{ib}.

$$BLER_{ib} = f(P_b^{Tx}, \boldsymbol{H_{ib}}, m_{ib}) \tag{3}$$

In Section IIIB, we elaborate how a look up table can be used in lieu of the function in (3). Henceforth, user QoS will refer to the user's throughput and BLER requirements.

B. BS Power Consumption Model

The RF chain consists of PA and RF chain transceiver circuitry. PA is the major contributor to BS power and has four states of operation namely, off, idle, active and switching states [18]. PA is switched off in the off state, and it is on but not transmitting in the idle state. PA transmits in the active state and the power consumption comprises of the idle power and transmission power. The transmission power consumption depends on PA efficiency, transmit power (assumed constant), bandwidth and duration of transmission. The switching power is comparable to idle power, however, the switching duration is much lower than time slot duration. Hence, the contribution of switching power is much lower than that of idle power when power consumption is averaged over the frame duration.

The baseband signal processing, DC-DC conversion, AC-DC conversion and cooling modules of the BS contribute significantly to BS power consumption. As they cannot be switched at the time scale of PA, the power consumption of the above modules has a baseline component independent of the PA state and an additional power component which scales with bandwidth of transmission when PA is transmitting. We adopt the model presented in [19] which captures the characteristics of BS module power consumption described above. The model in [19] is extended to include the off and switching power of PA and is briefly described below.

The frequency utilization ψ_{st} of RF chain $s \in [1, S]$ in time slot $t \in [1, T]$ due to $| I_b |$ users is

$$\psi_{st} = \begin{cases} \frac{1}{J} \sum_{i=1}^{|I_b|} J_{sti}, & \text{if PA is in active state} \\ 0, & \text{if PA is in idle or off state} \end{cases} \tag{4}$$

where J_{sti} is the number of frequency blocks assigned on RF chain $s \in [1, S]$ in time slot $t \in [1, T]$ to the i^{th} user. As in LTE systems, we consider frequency block allocation on a per time slot basis in a frame [20] to determine ψ_{st}. The number of active RF chains in a time slot t is $S_t^A =| \{s : \psi_{st} > 0\} |$. The number of active and idle time slots in a frame is given by $T^A =| \{t : S_t^A > 0\} |$, $T^I =| \{t : S_t^A = 0 \land \exists s \in [1, S] : s \text{ is on}\} |$. Denoting the duration of PA switching as t^{Sw} and the number of RF chains switching in a frame as S^{Sw}, the duration of all the RF chains in the off state in a frame is $t^O = t^F - \frac{t^F}{T}(T^A + T^I) - t^{Sw} S^{Sw}$.

Using the above definitions, the average power consumption of BS b with S RF chains in a frame with T time slots is

$$\begin{aligned} P_b = \frac{1}{t^F}(&\sum_{t=1}^{T_b^A}(S_{tb}^A P^I + \Delta_p P^{Max} \sum_{s=1}^{S_{tb}^A} \sum_{i=1}^{|I_b|} \psi_{stib} + \\ &(S - S_{tb}^A) P^O) + S T_b^I P^I) + S t_b^O P^O + S_b^{Sw} t_b^{Sw} P^{Sw} \end{aligned} \tag{5}$$

In the model above, P^O is the BS power consumption when the PA is switched off and includes the idle power consumption of all components excluding the PA and the off state power consumption of PA. The load independent term P^I represents the idle power of PA and the other components. The BS power consumption in the active time slots includes the baseline idle power component given by $S_{tb}^A P^I$ and the active power due to transmission modeled as the load dependent term $\Delta_p \psi_{st} P^{Max}$. The load dependent term $\Delta_p \psi_{st} P^{Max}$ increases linearly with only frequency utilization ψ_{st} as power gradient (slope) Δ_p and maximum transmit power P^{Max} are maintained constant. In the proposed technique, PA is either in the active, off or switching state. Henceforth, $T^I P^I$ is not a contributor to P_b. Defining $S_b^A = \{S_{tb}^A : t \in [1, T_b^A]\}$ and $\psi_b = \{\psi_{sti} : s \in [1, S_{tb}^A], t \in [1, T_b^A], i \in [1, | I_b |]\}$, the average cluster power consumption in a frame is given by

$$P_C = \sum_{b=1}^{|C|} P_b = f(\{(I_b, T_b^A, S_b^A, \psi_b) : b \in C\}) \tag{6}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

C. Problem Formulation

We can infer from (2-3, 5) that the QoS requirements and channel conditions of I_b users determine the aggregate BS resource utilization and P_b. At the individual BSs, given I_b, the BS resource space formed by number of RF chains S, time slots T and frequency blocks J can be explored during user mode selection to minimize P_b. At the cluster level, adapting the association of users $I_C = \cup_{b \in C} I_b$ will adapt the aggregate BS resource utilization and P_b. However, the association of all the users $I_b \forall b \in C$ cannot be adapted. This is because for every $b \in C$, there may exist a set of non-transferable users $I_b^{NT} \subseteq I_b$ that lie in the coverage area of only BS b and cannot be transferred to any other BS $b^{\sim} \in C \setminus b$ (see Fig. 2). The association of set of transferable users $I_b^T = I_b \setminus I_b^{NT}$ can be adapted as they lie in the coverage area of at least one more BS $b^{\sim} \in C \setminus b$ and can be transferred to BSs $\{b^{\sim}\}$. From the above description of I_b^{NT} and I_b^T, we can see that $I_b^{NT} \cap I_b^T = \emptyset \forall b \in C$. Further, assuming that a user is associated with no more than one BS, $I_b^T \cap I_{b^{\sim}} = \emptyset$ even though user $i \in I_b^T$ is located in the coverage area of BS $b^{\sim}$. Using the above, the set of cluster users is given $I_C = I_C^{NT} \cup I_C^T$ where $I_C^{NT} = \cup_{b \in C} I_b^{NT}$ and $I_C^T = \cup_{b \in C} I_b^T$ is the set of non-transferable and transferable cluster users respectively. The sets I_C^T and C together form the UA space that can be explored to adapt the set of users associated with BSs $b \in C$ and affect the individual BS resource utilization. The objective of the BS and UA resource adaptation is to maximize the number of RF chains that can be switched off in the cluster to minimize P_C while satisfying the QoS requirements in (2-3) for all the cluster users and not exceeding the BS resource utilization limits. The objective and constraints form the optimization problem stated below. Note, a single cluster C and associated users I_C is considered unless otherwise mentioned.

$$\min \sum_{b=1}^{|C|} \frac{1}{t^F} \Big(\sum_{t=1}^{T_b^A} \big(S_{tb}^A P^I + \Delta_p P^{Max} \sum_{s=1}^{S_{tb}^A} \sum_{i=1}^{|I_b|} \psi_{stib} + \tag{7}$$
$$(S - S_{tb}^A) P^O \big) \Big) + St_b^O P^O + S_b^{Sw} t_b^{Sw} P^{Sw}$$
$$\text{Subject to: } TP_{ib} \geq \gamma_i, \forall i \in I_C \tag{8}$$
$$BLER_{ib} \leq BLER_i^{Th}, \forall i \in I_C \tag{9}$$
$$\frac{t^F}{T} T_b^A + t^{Sw} S_b^{Sw} \leq t^F, \forall b \in C \tag{10}$$
$$S_{tb}^A \leq S, \forall t \in [1, T_b^A], \forall b \in C \tag{11}$$
$$\psi_{stb} \leq 1, \forall s \in [1, S_{tb}^A], \forall t \in [1, T_b^A], \forall b \in C \tag{12}$$

To minimize (7), the optimization variables are the sets $I_C^T = \cup_{b \in C} I_b^T$ and $\{T_b^A, \{S_{tb}^A\}, \{\psi_{stb}\} : b \in C, t \in [1, T_b^A], s \in [1, S_{tb}^A]\}$. The idle power and transmission power of the BS due to active RF chains (first and second terms in the summation over T_b^A in (7)) are the dominant components of P_b (Section IIB) and thereby, P_C. On the other hand, the off power due to inactive RF chains given by the third term in the summation over T_b^A is much lower than the static and dynamic powers and hence contributes less to the BS power consumption. This implies that the number of active RF chains will have priority in the optimization to minimize P_C. Minimizing the number of RF chains will result in minimizing the first and second terms of the summation over T_b^A while maximizing the third term in the summation over T_b^A. Further, minimizing the number of active RF chains in time slots to zero will maximize the RF chain off duration (t^O) and minimize the number of active time slots T_b^A. This will minimize the first term (entire summation over T_b^A) in (7) and maximize the second term (power consumption when all RF chains are off). Therefore, minimizing P_C can be considered equivalent to minimizing (maximizing) the number of active (off) chains. Constraints (8-9) respectively ensure that the throughput TP_{ib} and the $BLER_{ib}$ provided by BS b satisfies the i^{th} user's required rate γ_i and upper BLER bound $BLER_i^{Th}$. Constraint (10) ensures that the sum of duration of transmission and switching is upper bounded by t^F. The number of active RF chains in an active time slot is upper bounded by S in (11). The last constraint (12) specifies the upper bound on the frequency utilization of every active RF chain. An important point to note here is that satisfying the constraints (8-9) ensures that every cluster user is associated with a BS and therefore explicit constraints to ensure the same are not required. Henceforth, the optimization will be carried out with the transmission frame as reference.

III. Co-RFSnooze Algorithm

A. Multiple Multidimensional Knapsack Problem

The problem in (7-12) belongs to the class of Multiple Multidimensional Knapsack Problem (MMKP) as described below. Let the set of cluster users I_C and set of cluster BSs C denote the set of items and knapsacks respectively. UA is equivalent to assigning items to knapsacks and BS resource utilization is equivalent to utilizing the knapsack capacity. The profit of assigning user (item) $i \in I_C$ to BS $b \in C$ (knapsack) is the throughput TP_{ib} and the achievable $BLER_{ib}$ provided by BS b to user i. The number of BS RF chains S denotes the number of dimensions of the knapsack and the capacity of BS b in dimension $s \in [1, S]$ is JT, the total number of frequency blocks in a frame. The weight of user $i \in I_C$ in dimension $s \in S$ is the total number of frequency blocks assigned to the user in the frame given by $\sum_{t \in T} J_{sti}$. The BS resource and UA adaptation to minimize average cluster power consumption can be seen as MMKP with minimizing the total BS resource utilization, maximizing the users' throughput and minimizing the users' BLER as the criteria for optimization. The problem stated in (7-12) is a variant of the above multi-criteria MMKP which minimizes BS resource utilization subject to lower bound on throughput provided and upper bound on achieved BLER. As MMKP is a NP-Hard problem [21], we propose a heuristic algorithm that integrates BS resource and UA adaptation heuristics to solve (7-12).

B. BS Resource Adaptation - Heuristics and Algorithm

Consider the set of users I_b associated with BS b and let $I = | I_b |$. For brevity of notation, we will drop the subscript b in this subsection. Selection of mode $m_i \in M$ for the user $i \in I_b$ utilizes T_i active time slots, $s_{ti} \forall t \in [1, T_i]$ active RF chains and $J_{sti} \forall s \in [1, s_{ti}], t \in [1, T_i]$ frequency blocks. The mode selection for individual users impacts the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

overall BS utilization as follows.(i) $T^A = \max_{i=1,..,I} T_i$, (ii) $S_t^A = \max_{i=1,..,I} s_{ti}, \forall t \in [1, T_i]$ and (iii) $\psi_{st} = \sum_{i=1}^{I} \frac{J_{sti}}{J} \forall t \in [1, T_i], s \in [1, S_t^A]$. From the above, it can be inferred that T^A, S_t^A and ψ_{st} can be minimized if each is minimized for every user. However, minimizing each of the BS resource in isolation for every user will lead to an increase in the other BS resources because (a) decreasing T_i increases s_{ti} and J_{sti}, (b) decreasing s_{ti} increases T_i and J_{sti} and (c) decreasing J_{sti} increases T_i and s_{ti} in order to satisfy the QoS of the user. Therefore, joint adaptation of resources allocated to every user is required to minimize BS utilization and P_b.

The RFSnooze (Min-Cost in [1]) algorithm shown in Table II jointly adapts the BS resources to minimize BS utilization and P_b. The inputs to the algorithm are the required throughput γ_i and BLER threshold $BLER^{Th}$, the rank indicator RI_i and the channel quality indicator CQI_i sent as periodic feedback by all the users $i \in [1, I]$ [22], the channel matrix $\boldsymbol{H_i}$, the BS and user device resource upper bounds S, T, J and R. The steps of the algorithm are explained briefly below. The reader can refer to [1] for detailed explanation of the algorithm.

In step 4, the output of iterative frequency domain scheduler [23] is extended to allocate $T_i(m)$ time slots, $s_i(m)$ RF chains, $J_i(m)$ frequency blocks for all modes $m \in M$ in a frame for all users $i \in [1, I]$. The $BLER$ in step 5 is determined using the CQI and RI measurements and the Look Up Table (LUT) in [24] (used in lieu of BLER function in (3)) that specifies for different CQI values, the SINR threshold $SINR^{Th}(m)$ required for every mode $m \in M$ to result in $BLER \leq 0.1$. For all permissible modes $\{m : d_i(m) \leq RI_i\}$, if $SINR_i \geq SINR^{Th}(m)$ ($SINR_i$ is given by (1)), then $BLER_i(m) = BLER_i^{Th}$, else $BLER_i(m)$ is set to value greater than $BLER^{Th}$.

In step 6, the set of feasible modes $M_i^{FS} \subseteq M$ is updated with modes m that satisfies the throughput, BLER, and upper bounds on frequency and time utilization. From (5), the power consumption due to feasible mode $m \in M_i^{FS}$ is given by

$$P_i(m) = \frac{1}{t^F}(T_i(m)s_i(m)P^I + \frac{s_i(m)\Delta_p P^{Max}}{J}\sum_{t=1}^{T_i(m)} J_{ti}(m)) \tag{13}$$

The power consumption is calculated for every mode $m \in M_i^{FS}$ in step 7 and the mode m_i^* that results in minimum power consumption is chosen in step 8. The number of active time slots T^A, active RF chains $\{S_t^A : t \in [1, T^A]\}$, the frequency utilization $\{\psi_{st} : s \in [1, S_t^A], t \in [1, T^A]\}$ are the algorithm outputs determined in steps 10-14.

From Table II, the complexity of RFSnooze to determine the combination of modes is given by $| M | O(I)$ and is linear in I. In comparison, complexity of exhaustive search given by $O(| M |^I)$ is exponential in I.

C. UA Adaptation - Heuristics

SINR threshold for a mode m is defined as the threshold below which the BLER due to mode m, $BLER(m) > BLER^{Th}$ and can be determined as outlined in [24]. BS b that can provide SINR greater than the minimum of the SINR thresholds of all modes $m \in M$ can service the i^{th} user as there exists at least one mode m for which $SINR_{ib} > SINR^{Th}(m)$. Let E_i denote the set of BSs that can service the i^{th} user. We assume that the cluster users send the CQI and RI information for every BS $b \in C$ to the entire cluster [25]. Using this information, the BS-user assignment matrix $BSU = [k_{bi}]_{|C|x|I_C|}$ with elements $k_{bi} \in [0, | C |]$ is maintained at all BSs $b \in C$. The value $k_{bi} = 0$ indicates that BS $b \notin E_i$ as it does not satisfy the minimum of mode SINR thresholds for the i^{th} user. Sorting the BSs $b \in E_i$ in the decreasing order of SINR, the values $k_{bi} = 1$ indicates that BS b provides the highest SINR, $k_{bi} = 2$ indicates that BS b provides the second highest SINR to the i^{th} user and so on. Using the BSU matrix, the I_b^{NT} and I_b^T users associated with BS b can be defined as

$$I_b^{NT} = \{i : k_{bi} = 1 \wedge E_i = \{b\} \wedge \{v : k_{vi} \geq 2\} = \emptyset\} \tag{14}$$

$$I_b^T = \{i : k_{bi} = 1 \wedge E_i = b \cup \{v : k_{vi} \geq 2\}\} \tag{15}$$

Table III shows the BSU matrix for a cluster of size $| C |= 4$ and $| I_C |= 10$. Using (14-15), the sets I_b^{NT} and I_b^T for BSs $b = 1, 2, 3, 4$ can be written as: $I_1^{NT} = \{U3, U5\}$,$I_1^T = \{U7\}$;$I_2^{NT} = \emptyset$,$I_2^T = \{U1\}$;$I_3^{NT} = \{U4\}$, $I_3^T = \emptyset$;$I_4^{NT} = \{U8\}$,$I_4^T = \{U2, U6, U9, U10\}$. Note, for $BS2$, as $I_2^{NT} = \emptyset$ all the RF chains can be switched off by transferring $U1$. We will next discuss heuristics for allocating BS resources to I_b^{NT} and I_b^T users. Without loss of generality, we will consider BS $b \in C$ for the discussion and drop the subscript b for brevity.

From (5), the utilization of BS resources is the aggregate utilization due to $I^{NT} \cup I^T$. By allocating resources first to I^{NT} and subsequently to I^T, we can rewrite (5) as

$$\begin{aligned} P = \frac{1}{t^F}\Big(&\sum_{t=1}^{T^{NT}}(S_t^{NT}P^I + \frac{\Delta_p P^{Max}}{J}\sum_{s=1}^{S_t^{NT}}\sum_{i=1}^{|I^{NT}\cup I^{T^\sim}|} J_{sti}) \\ &+ \sum_{t=T^A-T^{NT}+1}^{T^A}((S_t^A - S_t^{NT})P^I + \frac{\Delta_p P^{Max}}{J}\sum_{s=S_t^A-S_t^{NT}+1}^{S_t^A} \\ &\sum_{i=1}^{|I^T\setminus I^{T^\sim}|} J_{sti}) + \sum_{t=1}^{T^A}(S - S_t^A)P^O\Big) + t^O S P^O + t^{Sw} S^{Sw} P^{Sw} \end{aligned} \tag{16}$$

where T^{NT} and S_t^{NT} are the number of active time slots and RF chains in time slot $t \in [1, T^{NT}]$ required to satisfy the QoS requirements of I^{NT} and $I^{T^\sim} \subseteq I^T$ users. This implies that $S_t^A - S_t^{NT}$ RF chains can be switched off in time slots $\{t \in [1, T^A] : S_t^A - S_t^{NT} > 0\}$ if $| I^T \setminus I^{T^\sim} |$ users are transferred to feasible cluster BSs. The subset of transferable users $I^{T^\sim}$ are updated as non-transferable users as their QoS requirements are satisfied by allocating no more than S_t^{NT} RF chains in time slots T^{NT} allocated to I^{NT} users. The possibility of reducing $| I^T |$ and complexity of UA is the motivation to allocate BS resources first to I^{NT} users and subsequently to I^T users. Next, we will select the "transferor" BS g which will transfer users and the "transferee" BSs E to transfer users to.

Higher the number of RF chains $S_t^A - S_t^{NT}$ that can be switched off, higher the savings in transferor BS power consumption. However, as the number of users $| I^T \setminus I^{T^\sim} |$ that are transferred increases, the number of users that receive less than maximum SINR and the transferee BS power consumption also increases. To maximize $S_t^A - S_t^{NT}$ while

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

TABLE II
RFSNOOZE ALGORITHM

Input: I_b, $\{\gamma_i, BLER_i^{Th}, RI_i, CQI_i, \boldsymbol{H_i} : i \in [1, I]\}, S, J, R, T$
Output: T^A, $\{S_t^A : t \in [1, T^A]\}$, $\{\psi_{st} : s \in [1, S_t^A], t \in [1, T^A]\}$

1. For all users $i \in [1, I]$
2: Initialize $M_i^{FS} = \emptyset$, $J_i(m) = 0$, $T_i(m) = 0$, $\forall m \in M$
3: For all modes $m \in M$
4: Scheduler updates $T_i(m) = max_{t\in[1,T]}\{t : J_{ti} > 0, J_i(m) = \sum_{t=1}^{T_i(m)} J_{ti}(m)$ if $TP_i(m, J_i(m), T_i(m))) \geq \gamma_i$
5: Determine $BLER_i(P^{Tx}, \boldsymbol{H_i}, m)$ using CQI_i entry in LUT
6: If $BLER_i(P^{Tx}, \boldsymbol{H_i}, m) \leq BLER_i^{Th}, d_i(m) \leq RI_i(m), T_i(m) \leq T, J_{ti} \leq JT$, then update $M_i^{FS} = M_i^{FS} \cup m$
7: Compute $P_i(m)$ using (13)

8: Find mode $m^* = argmin_{m\in M_i^{FS}} P_i(m)$
9: Update $T_i = T_i(m_i^*)$, $s_{ti} = s(m_i^*)$, $\psi_{sti} = J^{-1} J_{ti}(m_i^*)$, $\forall s \in [1, s_{ti}]$, $\forall t \in [1, T_i]$
10: Determine $T^A = max_{i\in[1,I]} T_i$
11: For all time slots $t = 1, .., T^A$
12: Determine $S_t^A = max_{i\in[1,I]} s_{ti}$
13: Determine off RF chains $S_t^O = S - S_t^A$
14: Determine $\psi_{st} = J^{-1} \sum_{i=1}^{I} J_{sti}$, $\forall s \in [1, S_t^A]$; $\psi_{st} = 0$, $\forall s \in [1, S_t^O]$

TABLE III
ILLUSTRATION OF BSU MATRIX WITH $\mid C \mid = 4$ AND $\mid I_C \mid = 10$

BS-User	1	2	3	4	5	6	7	8	9	10
1	0	4	1	0	1	3	1	0	2	0
2	1	3	0	0	0	4	2	0	3	2
3	2	2	0	1	0	2	3	0	0	0
4	0	1	0	0	0	1	4	1	1	1
Modified BSU matrix after restricting $E_i = \{b : k_{bi} \in [1, 2]\}$										
1	0	0	1	0	1	0	1	0	2	0
2	1	0	0	0	0	0	2	0	0	2
3	2	2	0	1	0	2	0	0	0	0
4	0	1	0	0	0	1	0	1	1	1

minimizing $\mid I^T \setminus I^{T\sim} \mid$ and the increase in transferee BS power consumption, the RF chain-user ratio RFU is defined as

$$RFU = \frac{\sum_{t=T^A-T^{NT}+1}^{T^A} S_t^A - S_t^{NT}}{\mid I^T \setminus I^{T\sim} \mid} \tag{17}$$

Larger the RFU ratio, higher will be the savings in transferor BS power consumption and lower will be the number of users receiving less than maximum SINR. Also, large RFU ratio will result in lower increase in transferee BS power consumption. Hence, the BS with the largest RFU ratio is nominated as the transferor BS g. Amongst the multiple BSs which cover user $i \in I_g^T \setminus I_g^{T\sim}$, the selection of transferee BS is restricted to that subset of BSs $b \in E_i$ with $k_{bi} = 2$ in the BSU matrix. This has a two-fold effect of reducing (a) the impact on QoS of the user $i \in I_g^T \setminus I_g^{T\sim}$ and (b) the complexity of UA. The set of transferee BSs corresponding to $I_g^T \setminus I_g^{T\sim}$ is denoted as E.

The above selection criterion is applied to Table III resulting in replacing all the entries with $k_{bi} > 2$ with $k_{bi} = 0$ to indicate that BS b is not a transferee BS for the i^{th} user. The bottom portion of Table III shows the modified BSU matrix. This reduces $\mid E_i \mid$ for i^{th} user and also minimizes the impact on the user QoS. For instance the set of transferee BSs for $U7$ is reduced from $E_7 = \{BS1, BS2, BS3, BS4\}$ to $E_7 = \{BS1, BS2\}$.

We will now discuss the three feasibility conditions that have to be satisfied for transferring users. The first condition is that the QoS requirements of transferrable users of transferor BS and the users of transferee BS have to be satisfied by the transferee BS after the transfer.

$$C1 : \text{satisfy constraints (8-9)} \forall e \in E, i \in I_e \cup I_g^T \setminus I_g^{T\sim} \tag{18}$$

Let us denote the number of active time slots, active RF chains and frequency utilization of BS b before user transfer as T_b^A, S_b^A, ψ_b and after user transfer as $T_b^{A*}, S_b^{A*}, \psi_b^*$. The second condition is that BS resource utilization of transferee BS e after transfer $T_e^{A*}, S_e^{A*}, \psi_e^*$ should satisfy (10-12).

$$C2 : \text{satisfy constraints (10-12)} \forall e \in E \tag{19}$$

Denoting the power consumption of BSs after user transfer as P^*, the third condition is that the difference in cluster power consumption before and after transfer should be positive.

$$C3 : \Big(P_g(I_g, T_g^A, S_g^A, \psi_g) + \sum_{e=1}^{|E|} P_e(I_e, T_e^A, S_e^A, \psi_e) - P_g^*(I_g^{NT} \cup I_g^{T\sim}, T_g^{A*}, S_g^{A*}, \psi_g^*) - \sum_{e=1}^{|E|} P_e^*(I_e^{NT} \cup (I_g^T \setminus I_g^{T\sim}), I_e^T, T_e^{A*}, S_e^{A*}, \psi_e^*) \Big) > 0 \tag{20}$$

D. Co-RFSnooze Algorithm

The Co-RFSnooze algorithm adopts a bottom-up iterative approach which adapts BS resources at individual cluster BSs and adapts UA at cluster level in an iterative manner. An iteration consists of two key interlinked steps explained below. The first key step is that the Co-RFSnooze algorithm applies the RFSnooze algorithm at each cluster BS to I^{NT} and subsequently to I^T users and determines the RFU ratio. This step (a) minimizes the number of RF chains required to satisfy the QoS requirements of I^{NT} users at the individual BS level, (b) reduces the cardinality of the I^T (Section IIIC) to prune the UA space at the cluster level and (c) determines the BS resources required to satisfy the QoS requirements of the $I^T \setminus I^{T\sim}$ users using which the RFU ratio is calculated. The RFU ratio guides the choice of transferor BS and is the crucial link between individual BS resource adaptation and cluster level UA adaptation. The second key step is the selection of transferor and transferee BSs. The BS with highest RFU ratio is selected as the transferor BS to maximize the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

TABLE IV
CO-RFSNOOZE ALGORITHM

Input: $\{I_b^{NT}, I_b^T : b \in [1, \mid C \mid]\}, \{\gamma_i, BLER_i^{Th} : i \in [1, \mid I_C \mid]\}, \{RI_{ib}, CQI_{ib}, \boldsymbol{H_{ib}} : i \in [1, \mid I_C \mid], b \in [i, \mid C \mid]\}, S, J, R, T$
Output: $\{I_b, T_b^A, \{S_{tb}^A\}, \{\psi_{stb}\} : s \in [1, S_{tb}^A], t \in [1, T_b^A], b \in [1, \mid C \mid]\}$
1. Initialize set of possible transferor BSs $G = C$, set of transferee BSs $E = \{\}$, transferor BS $g = \{\}$
2. For all BSs $b \in C$ 3: Initialize I_b^{NT} and I_b^T using (14) and (15) 4: Apply RFSnooze to I_b^{NT} to determine BS resource allocation for I_b^{NT} 5: Apply RFSnooze to I_b^T to determine BS resource allocation for I_b^T 6: Determine $I_b^{T\sim} \subseteq I_b^T$ that require no additional time slots and RF chains as compared to I_b^{NT} 7: Update $I_b^{NT} = I_b^{NT} \cup I_b^{T\sim}, I_b^T = I_b^T \setminus I_b^{T\sim}$, update BSU_b with $k_{ei} = 0, \forall e \in C \setminus b$ 8: Calculate P_b using (5) and RFU_b using (17)
9: If $G = \{\}$, then go to step 27, Else 10: Select transferor BS with highest RFU ratio $g = max_{b \in G} RFU_b$ 11: Update $G = G \setminus g$ 12: Determine subset of BSs $E = \{e : \exists i \in I_g^T \setminus I_g^{T\sim} \wedge k_{ei} = 2\}$ to which BS g can transfer users $I_g^T \setminus I_g^{T\sim}$
13:For all BSs $e \in E$ 14: Update $I_e^{NT} = I_e^{NT} \cup \{i : i \in I_g^T \setminus I_g^{T\sim} \wedge k_{ei} = 2\}$ 15: Apply RFSnooze to I_e^{NT} to determine BS resource allocation for I_e^{NT} 16: Apply RFSnooze to I_e^T to determine BS resource allocation for I_e^T 17: Determine P_e^* using (5) and $\Delta P_e = P_e - P_e^*$ 18: If transfer feasibility condition $C1$ or $C2$ is violated 19: Then set $P_e^* = \infty, \Delta P_e = \infty$
20: Apply RFSnooze to I_g^{NT} users of transferor BS g to determine BS resource allocation 21: Determine P_g^* using (5) and $\Delta P_g = P_g - P_g^*$ 22: If transfer feasibility condition $C3$ is true, then for all users $i \in I_g^T \setminus I_g^{T\sim}$, for all BSs $e \in E$ 23: Update the BSU matrix $k_{gi} = 0, k_{ei} = 1$ 24: Else for all users $i \in I_g^T \setminus I_g^{T\sim}$, for all BSs $e \in E$ 25: Update the BSU matrix $k_{ei} = 0$ 26: Go to step 2
27: For all BSs $b \in C$ 28: $I_b = \{i : k_{bi} = 1\}, \{T_b^A, \{S_{tb}^A\}, \{\psi_{stb}\} : s \in [1, S_{tb}^A], t \in [1, T_b^A]\}$ - Output of step 4

savings in power consumption due to switching off RF chains and minimize the impact on users' received SINR. The set of transferee BSs is restricted to BSs that provide the second highest SINR to $I^T \setminus I^{T\sim}$ of transferor BS to reduce UA space. The above two key steps are carried out iteratively by Co-RFSnooze algorithm as described below.

The Co-RFSnooze algorithm is shown in Table IV. The algorithm inputs are the set of cluster users, their QoS requirements and the channel state information, the BS resource upper bounds for the cluster BSs. The algorithm outputs are the set of users associated with each of the cluster BSs and corresponding resource utilization of the BS.

Starting with the set of transferor BSs $G = C$ and set of transferee BSs $E = \emptyset$, the algorithm iterates till the set of transferor BSs $G = \emptyset$. Each iteration starts by allocating individual BS resources first to I_b^{NT} users in step 4 and subsequently to I_b^T users in step 5. The set of users $I_b^{T\sim}$ that can be serviced in T_b^{NT} time slots with $S_{tb}^{NT}, t \in [1, T_b^{NT}]$ RF chains is obtained from step 6. The sets I_b^{NT} and I_b^T are updated in step 7 and the power consumption P_b and the RFU ratio are calculated in step 8.

Using the RFU ratio, steps 10-11 selects the transferor BS g and updates the set of transferor BSs G to exclude the selected BS g. The set of transferee BSs E is selected in step 12 and the corresponding sets of $I_e^{NT}, \forall e \in E$ are updated in step 14 to include the transferable users $I_g^T \setminus I_g^{T\sim}$ of BS g. The update of G and of $I_e^{NT} \forall e \in E$ is of particular importance. By updating the set $G = G \setminus g$ in the current iteration eliminates the selection of BS g as transferor BS in any subsequent iterations. This reduces the cardinality of set of possible transferor BSs G for subsequent iterations and ensures convergence of the algorithm in at most $\mid C \mid$ iterations. The update $I_e^{NT} = I_e^{NT} \cup I_g^T \setminus I_g^{T\sim}$ categorizes $I_g^T \setminus I_g^{T\sim}$ of BS g as non-transferable users of BS e. This will not allow oscillatory behavior wherein the users $I_g^T \setminus I_g^{T\sim}$ are assigned back to the transferor BS g in subsequent iterations in which transferee BS e may be selected as transferor BS and BS g as transferee BS.

The BS resource allocation taking in to account the transferred users is determined in steps 15-16 following which the transfer feasibility conditions C1, C2 and C3 (Section IIIC) are tested in steps 18-22. Note that condition C1 is implicitly satisfied by the RFSnooze algorithm as it selects feasible modes which satisfies the constraints (8-9) for each user. Iterative allocation of resources to users as explained in Section IIIB, [1] ensures that the BS resource utilization constraints (10 -12) are satisfied. Given the resource utilization of BSs g and E, C3 is evaluated using (20). If conditions C1, C2 and C3 hold, then the BSU matrix entries for users $I_g^T \setminus I_g^{T\sim}$ are updated in step 23 to reflect the disassociation from transferor BS g ($k_{gi} = 1$ to $k_{gi} = 0$) and association with the transferee BS e ($k_{ei} = 2$ to $k_{ei} = 1$). If the conditions do not hold, then the BSU matrix is updated in step 25 to reflect that the users $I_g^T \setminus I_g^{T\sim}$ are non-transferable users of BS g ($k_{ei} = 2$ to $k_{ei} = 0$). In addition the power consumption of all transferee BSs is set to an arbitrarily large number to indicate that the transfer is not feasible. This is carried out for implementation purposes as elaborated in the next subsection. With the updated UA and set of possible transferor BSs G, the next iteration is initiated in step 26.

The iterations terminate when there are no more candidates for transferring users, i.e., $G = \emptyset$. In the final iteration, steps 2-8 are executed, however, since there are no more transferable users, the BS resource allocation obtained in step 4 is the final BS resource allocation. The check in step 9 is true for the final iteration and the algorithm terminates by executing steps 27-28. The outputs of the algorithm are the UA obtained from the BSU matrix and the corresponding BS resource utilization of the cluster BSs. We will use the example in Table III (bottom portion) with cluster of size $\mid C \mid= 4$ and $\mid I_C \mid= 10$ users to run through the algorithm steps with the aid of Fig. 3. The rows of Fig. 3 illustrate the BS resource utilization for each BS at the beginning of an iteration and lists the subsequent steps. The BS resource utilization is shown for one time slot of a transmission frame with $J = 24$ frequency blocks available on each of $S = 4$ RF chains $(S_1, .., S_4)$. The maximum number of user RF chains is $R = 4$. The frequency blocks allocated to users are indicated by the color used for the user. Due to lack of space, we have omitted showing multiple time slots in the transmission frame. For each user, the modes $m \in M_i^{FS}$ and the corresponding allocation of time slots and frequency blocks are listed in the legend using a 5-tuple - $(s_i, r_i, d_i, J_i, T_i)$. The I^T of each BS are differentiated by two vertical black colored lines placed on the BS resources allocated. For instance, $I^T = \{U7\}$ for $BS1$ and two black lines are placed on the yellow blocks on S_1 RF chain.

Initially $G = \{BS1, BS2, BS3, BS4\}$, $E = \emptyset$. The top row of Fig. 3 shows the set of feasible modes M^{FS} (Section IIIB) and the minimum power mode $m*$ (indicated by the tick mark) selected for I^{NT} and I^T of BSs $BS1, BS2, BS3, BS4$ in steps 4 and 5 of iteration 1. The outputs of steps 1-28 for iteration 1 are listed below the BS resource utilization illustration. At the end of iteration 1, the RF chain requirements at $BS1 = \{S_1, S_2, S_3, S_4\}$, $BS2 = \emptyset$, $BS3 = \{S_1, S_2\}$ and $BS4 = \{S_1, S_2, S_3, S_4\}$. Due to transfer of $U1$ from $BS2$ to $BS3$, 2 RF chains are switched off at $BS2$ in iteration 1. This is the initial BS resource utilization of iteration 2 shown in second row of Fig. 3. The steps 4-26 of iteration 2 result in transfer of $U2, U9$ from $BS4$ to BSs $BS1, BS3$ and switching off RF chains S_2, S_3, S_4 of $BS4$. This is shown in the third row of Fig. 3. The algorithm terminates with the third iteration as RFU ratios $RFU_1 = 0, RFU_2 = 0, RFU_3 = 0, RFU_4 = 0$. We can see that Co-RFSnooze reduces the number of active RF chains from 12 to 7 in the cluster by iteratively applying the RFSnooze algorithm and UA adaptation heuristics.

E. Complexity Analysis

As exhaustive search of UA space evaluates $\mid C \mid^{|I_C^T|}$ combinations, the complexity of UA adaptation is $O(\mid C \mid^{|I_C^T|})$. For each UA combination, the exhaustive search of the BS resource space has to evaluate $\mid M \mid^{|I_1|} +..+ \mid M \mid^{|I_{|C|}|}$ combinations. Therefore, the complexity of joint search of BS resource spaces and US spaces is given by $O(\mid C \mid^{|I_C^T|} (\mid M \mid^{|I_1|} +..+ \mid M \mid^{|I_{|C|}|}))$. The Co-RFSnooze algorithm evaluates a single combination of UA in an iteration and the maximum number of iterations for convergence of Co-RFSnooze is $\mid C \mid$. The complexity of UA space search is $O(\mid C \mid)$. In each iteration, the RFSnooze algorithm is executed at most twice for the entire cluster (steps 4-5, 15-16 and 20 in Table IV). The number of operations when RFSnooze algorithm (Section IIIB) applied to the every BS of entire cluster is $\sum_{b=1}^{|C|} \mid M \mid\mid I_b \mid=\mid M \mid\mid I_C \mid$. The complexity of the Co-RFSnooze algorithm for determining the BS resource allocation and UA in $\mid C \mid$ iterations is given by $2 \mid C \mid\mid M \mid O(\mid I_C \mid)$ where $\mid C \mid$ and $\mid M \mid$ are constants for a given cluster and BS resource configurations. Hence, Co-RFSnooze algorithm achieves linear complexity compared to the exponential complexity of exhaustive search.

F. Co-RFSnooze Framework

We propose a combination of the centralized approach [26] and the decentralized approach in [25] for the Co-RFSnooze framework to minimize the exchange of user QoS, channel state information (CSI) and control information between the cluster BSs to adapt UA.

The cluster BSs send training sequences to all the cluster users periodically [22]. In response, as implemented in decentralized approach in [25], the users estimate the CSI for each of the BS in the cluster and then send $\mid C \mid$ CSI estimates as feedback to every BS in the cluster. In this manner, the cluster BSs have the information about the SINR received by i^{th} user from every cluster BS $b \in C$. This enables the BSs to build and maintain a copy of the BSU matrix locally denoted as BSU_b. With the aid of Table IV and Fig. 4, we will next discuss information exchange required for the Co-RFSnooze iterations.

With the inputs required and BSU matrix available at the BSs, steps 2-7 (Table IV) are run at every BS $b \in C$ for updating I^T. Subsequently, the BSs broadcast their RFU values to all the other cluster BSs. The BS with highest RFU ratio selects itself as the transferor BS with the other BSs implicitly getting this information from the broadcasted RFU values. Using the updated local copy of BSU matrix, the transferor BS g determines the set of transferee BSs E as in step 12. The above operations are listed in boxes in Fig. 4.

We adopt the cooperation protocol in [26] to set up the communication interface between BS g and BSs $e \in E$ shown in Fig. 4. The BS g sends the "Transferor Request" to BSs $e \in E$ which in turn sends the "Transferee Ack" response to complete the cooperation setup. The BS g transmits to each BS $e \in E$, the row $k_{e*} \in BSU_g$ corresponding to BS e. Note that the row $k_{e*} \in BSU_g$ transmitted by BS g is identical to the row $k_{e*} \in BSU_e$ (local copy of BSU matrix at BS e) except for the entries corresponding to $i \in I_g^{T\sim}$ for which $k_{ei} = 0, k_{ei} \in BSU_g$ (as updated in step 7, Table IV) and $k_{ei} = 2, k_{ei} \in BSU_e$. This difference indicates to BS e the reduced set of users $I_g^T \setminus I_g^{T\sim}$ required for steps 13-19. The QoS requirements $(\gamma_i, BLER_i)$ of the users $\{i : i \in I_g^T \setminus I_g^{T\sim}\}$ required as input to RFSnooze algorithm in steps 15-16 are transmitted to the transferee BS. Execution of RFSnooze algorithm in steps 15-16 will implicitly evaluate conditions C1 and C2, which if violated will set the difference power consumption ΔP_e to an arbitrarily large value. The ΔP_e is conveyed to BS g by all BSs $e \in E$ which evaluates condition C3. The BSU_g

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

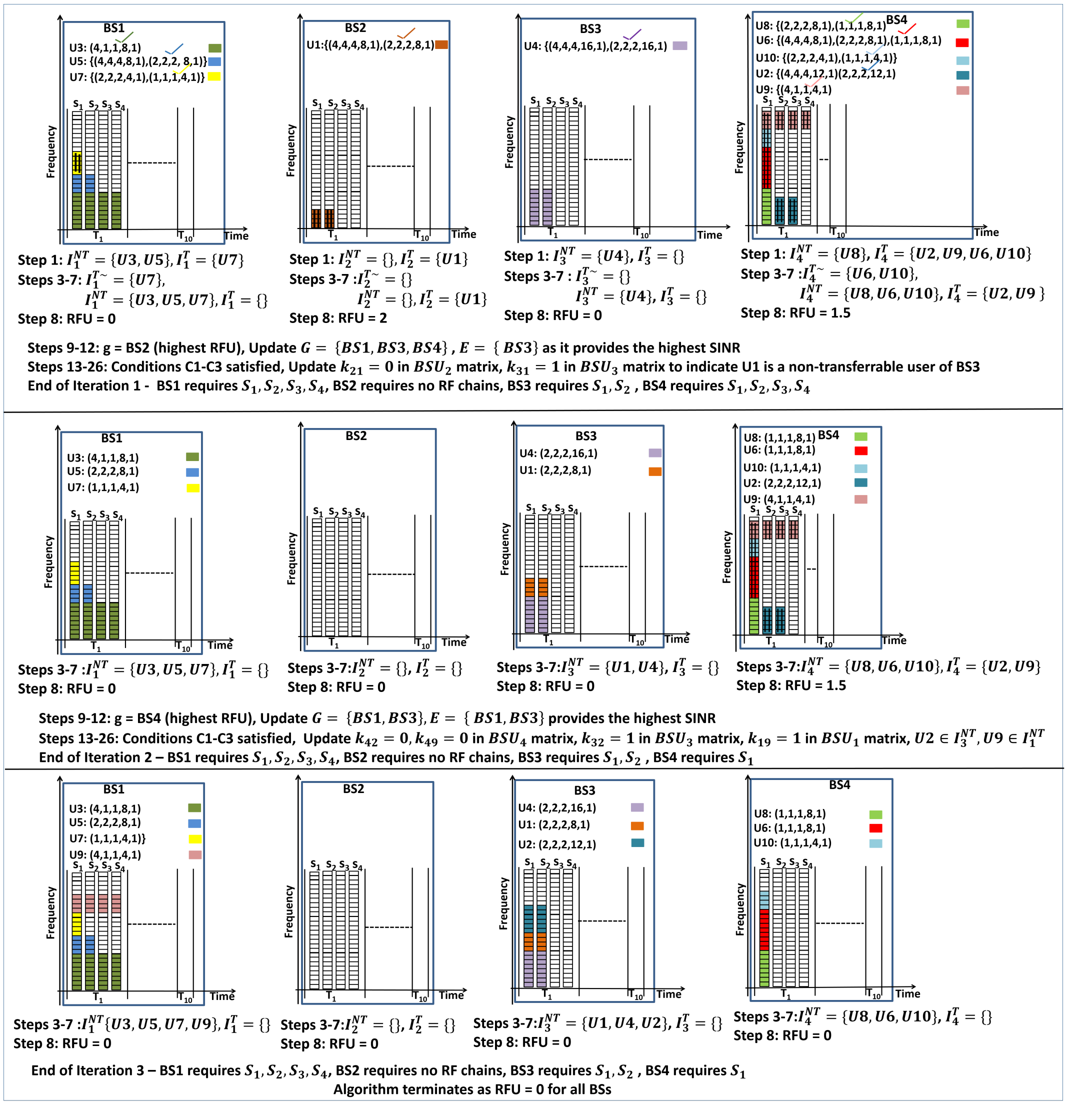

Fig. 3. Application of Co-RFSnooze algorithm to example in Table III

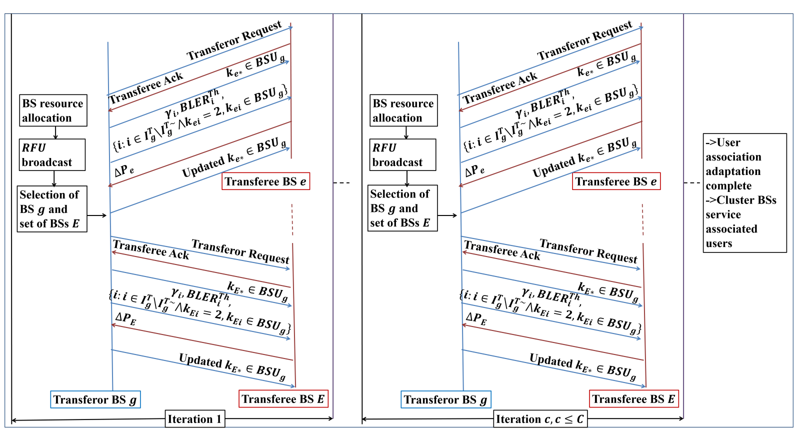

Fig. 4. Implementation of Co-RFSnooze Algorithm

TABLE V
SIMULATION PARAMETERS

Power gradient Δ_p	4.2
Off power P^O, Idle Power P^I	82.75W, 186W
PA switching power P^{Sw}, switching time t^{Sw}	100W, 35us
Maximum transmit power P^{Max}	40W
Bandwidth BW, Number of frequency blocks J	20MHz, 100
Duration of frame t^F, Number of time slots T	10ms, 10
Number of RF chains at BS S and user device R	4, 4
Set of modes M, $\mid M \mid$	{(1,1,1) (SISO), (2,2,2) (SM), (2,2,1) (SD), (4,1,1) (SD), (4,4,4) (SM), (4,2,2) (SM-SD)}, 6
Size of cluster $\mid C \mid$	4
Maximum number of cluster users	300
$BLER^{Th}$ for all cluster users	0.1
Simulation time	24 hours

matrix is updated as per step 23 or step 25 depending on evaluation of condition C3. The updated rows $k_{e*} \in BSU_g$ are transmitted to BSs $e \in E$ and the current iteration ends. The c^{th} iteration consists of the operations indicated by the boxes and information exchange shown in Fig. 4. After a cluster BS has been selected as transferor BS, in subsequent iterations, it broadcasts $RFU = 0$ value. In terms of implementation, when all the BSs broadcast $RFU = 0$, the algorithm terminates. Subsequently, the cluster BSs use the updated local BSU matrices to service the associated users.

The overhead due to information exchange among the cluster BSs is as follows. A byte each for mantissa and exponent is sufficient to represent RFU values. The size of BSU row given by $\lceil(\log_2 \mid C \mid)\rceil \mid I_C \mid$ depends on the cluster size and number of cluster users. Two bytes are sufficient to convey the QoS requirements of each of the users $i \in I_g^T \setminus I_g^{T\sim}$. The ΔP_e values can be expressed using a byte each for mantissa and exponent. Analysis in [17] shows that the gains due to adding a BS to the cluster significantly decreases when $\mid C \mid > 4$. Assuming $\mid C \mid = 4$ and $\mid I_C \mid = 300$, the BSU row, RFU byte, ΔP_e value and QoS information will account for $600 + 8 + 16 + 16* \mid I_g^T \setminus I_g^{T\sim} \mid$ bits. Assuming 0.5uW [13] is consumed for every bit transmitted over the backhaul, number of iterations is $\mid C \mid = 4$ and total number of users transferred $\mid I_g^T \setminus I_g^{T\sim} \mid = 35$ (Fig. 6b, high load), then the overhead due to information exchange for Co-RFSnooze is 2.368mW. Note that the overhead due to information exchange in iterations has been accounted in the calculation of P_C for the Co-RFSnooze algorithm in Section IVB.

The time scale of BS resource allocation is of the order of milliseconds as current LTE standards allows BS resource allocation every time slot (1ms duration) in a transmission frame. UA adaptation requires user transfer/handover from the transferor BS to the transferee BS. In this paper, it is assumed that the cluster BSs are connected via X2 interface and X2 handovers can be used to achieve the user transfer. Experiments in [27] show that the X2 handovers can take up to 100ms. Therefore, the time required for BS resource adaptation is about f times ($f = 10$ with the values considered) lesser than that required for UA adaptation and results in a two time scale system. The Co-RFSnooze algorithm accomodates the two time scale requirement as follows. Steps 4-5 in Table IV are carried out at periodicity of p_{BR} at individual BSs to adapt BS resource utilization. At periodicity $f * p_{BR} > p_{BR}$, all the iterations of the algorithm executing all the steps in Table IV are carried out to determine the BS resource allocation and UA of cluster BSs. In Section IVB, we evaluate the performance of Co-RFSnooze algorithm at a single time scale using the sample load trace from anonymous operator with granularity of 1 minute. We have chosen a single time scale of 1 minute ($f * p_{BR}$) as it satisfies the time scale requirements of both the adaptations as well reduces the overhead due to user transfer and allows evaluation of the Co-RFSnooze performance in its entirety, i.e, execute all the iterations at every point of the trace. Note, however, the evaluation can be easily extended to show the two time scale operation of Co-RFSnooze.

IV. SIMULATION FRAMEWORK AND RESULTS

A. Simulation Framework

In this section, we describe the simulation framework developed and the simulation parameters listed in Table V. We adopt the topology with 15 BSs in $4.5x4.5km^2$ [28], a part of 3G network in urban environment. The inter-cell distance is 0.5km. The cluster size $\mid C \mid$ is set to 4 and a 16^{th} BS is randomly placed in the considered 15 BS topology to obtain 4 clusters. Without loss of generality, we consider one of the four clusters to evaluate the proposed Co-RFSnooze algorithm. The BS power model presented in Section IIB is used to estimate the average BS and cluster power consumption in a frame. The BS power consumption parameters are specified in [19] and [18] and listed in Table V. The users (maximum 300) are uniformly and randomly distributed in the cluster. The traffic load is assumed to be spatially heterogeneous with user's required rate $\gamma \propto (\max(d) - d^2)$ where d is the distance between the user and BS. The $BLER$ LUT table in [24] is extended to include the modes (4,4,1) and (4,4,4) and used to determine the $BLER$ of users as explained in Section IIIB. Other parameters for the simulations follow the suggestions in the LTE specifications [20]. We consider the COST-231 HATA model for the path loss between the BS and user [29].

For comparing the performance of Co-RFSnooze algorithm, we consider the following algorithm/schemes (Section IA):

- All-On (conventional scheme): turns on all BS RF chains in active time slots and turns off in off slots.
- RFSnooze [1]: adapts number of active RF chains, time slots and frequency blocks at individual BSs in an uncoordinated manner. RFSnooze [1] has been extended to Co-RFSnooze algorithm in this paper.
- Co-Nap [11]: adapts the on/off pattern of the cluster BSs and turns off all BS RF chains to switch off BSs. The short time scale operation of BS switching effected by switching on/off all RF chains in a cooperative manner without using CoMP transmission makes Co-Nap the most relevant prior art technique for comparison.
- Exhaustive search: yields the combination that switches off the optimal number of RF chains

We will now discuss the implementation details of All-On and Co-Nap. The UA rule for All-On and Co-Nap schemes is that the user is associated with that BS which provides the highest SINR. The scheduling algorithm [23] (Section IIIB, [1]) is used to determine the feasible set of modes M^{FS}. As all the RF chains are switched on during the active time slots for All-On and Co-Nap, the mode that utilizes all the RF chains and satisfies the minimum throughput and BLER constraints is selected from the feasible mode set. If the QoS constraints are not satisfied by modes utilizing all the RF chains, then the mode with next highest number of RF chains that satisfies the QoS constraints is selected. The dominant operation in mode selection is determination of M^{FS} and is carried out as explained in Section IIIB, [1] for All-On, Co-Nap and RFSnooze. Hence, the the complexity of mode selection for All-On and Co-Nap is given by $\mid M \mid O(\mid I_C \mid)$ (Section IIIB). In case of All-On and Co-Nap, RF chains that are not transmitting in active time slots (in a frame) are in the idle state and by the UA rule, the set $I_b^T = \emptyset, I_b = I_b^{NT} \forall b \in C$. Incorporating the above in to (5), the BS average power consumption in a frame is

$$P = \frac{1}{t^F}\left(\sum_{t=1}^{T^A} SP^I + \frac{\Delta_p P^{Max}}{J}\sum_{s=1}^{S_t^A}\sum_{i=1}^{|I^{NT}|} J_{sti}\right) + t^O SP^O \qquad (21)$$

All-On does not adapt switching of BSs and RF chains. In contrast, Co-Nap adaptively switches on/off BSs and impacts the average power consumption of the cluster as briefly explained below. Co-Nap divides the transmission time into discrete transmission cycles comprising of $\mid C \mid$ number of blocks. The BS on/off (flickering) pattern determines the active and inactive (napping) blocks for all the BSs in every transmission cycle. The BS resource allocation is carried out for all the active blocks in a manner that the user QoS requirements are satisfied. Assuming that a block spans over multiple frames, P_b in a frame in an active time block is given by (21). For a frame in an inactive block (BS off), (21) reduces to SP^O (as $t^O = t^F$). For Co-Nap, the complexity of determining the on/off (1/0) pattern for $\mid C \mid$ BSs in $\mid C \mid$ blocks and BS resource allocation for $\mid I_C \mid$ cluster users is given by $\mid C \mid O(2^{|C|}) + \mid M \mid O(\mid I_C \mid)$.

B. Simulation Results

We will now present the experimental results obtained using the simulation framework described above. In order to evaluate the performance of the comparison schemes and the proposed algorithm in a practical setting, we adopt the sample traffic trace shown in Fig. 5a. The sample traffic trace is the normalized BS utilization measured by an anonymous operator in [30] for 24 hours with granularity of 1 minute. The simulation step is fixed as 1 minute, however, our simulation framework supports simulation step lesser than or greater than 1 minute. Fig. 5b shows the number of users in a simulation step. It is given by the product of value of the sample trace and maximum number of cluster users (Table V). Assuming that the number of users and their requirements do not change over the simulation step, the comparison schemes/algorithms and Co-RFSnooze algorithm is run once in every simulation step to

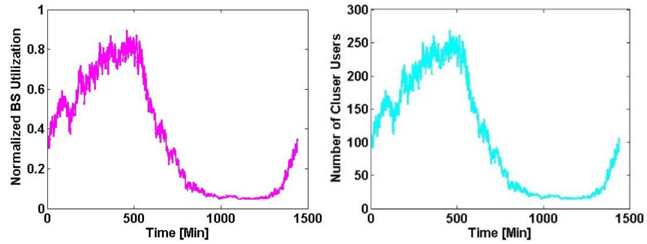

Fig. 5. Sample traffic trace, (b) Number of cluster users

determine the BS resource allocation for all the frames and in case of Co-RFSnooze, additionally, the updated UA. The P_C in a simulation step is the power consumption averaged over all the frames in a simulation step and is estimated using (6) for the proposed algorithms and using (21) in (6) for All-On. For Co-Nap, the simulation step is equivalent to the transmission cycle and consists of $\mid C \mid= 4$ blocks of equal duration. Co-Nap is run once every simulation step to determine the number of active blocks and resource allocation for all the frames in the active blocks. The P_C in a simulation step is equal to the power consumption averaged over the four blocks.

Fig. 6a shows the average power consumption of the cluster in a frame P_C for All-On (shown in red), RFSnooze (shown in blue) and Co-RFSnooze (shown in green). All-On consumes higher power than proposed algorithms because, regardless of the load, all the RF chains are on in the active time slots. This increases total RF chain power consumption due to (a) frequency utilization of each active RF chain and (b) idle power of the RF chain transceiver circuitry as all RF chains are either in active or idle state. Joint adaptation of number of active RF chains, frequency and time utilization reduces the cluster power consumption for RFSnooze. The green plot in Fig. 6a shows that the savings due to RFSnooze is further extended by Co-RFSnooze. This increase in power savings validates our extension of RFSnooze to Co-RFSnooze which, as elaborated in Section IIID, integrates BS resource adaptation and UA to maximize the number of cluster RF chains that can be switched off. Under high load conditions, RFSnooze achieves up to 35% gains (635^{th} minute) and Co-RFSnooze achieves up to 56% gains (382^{nd} minute) compared to All-On. RFSnooze achieves up to 42% gains (1151^{th} minute) and Co-RFSnooze achieves 49% gains (960^{th} minute) compared to All-On under low load conditions. Note that we refer to the savings in average cluster power consumption as the gains achieved.

We will now compare the performance of RFSnooze and Co-RFSnooze using Figs. 6a and 6b. Fig. 6b shows the number of users transferred by Co-RFSnooze during UA adaptation. Under high load conditions, Fig. 6b shows that higher number of users is transferred (up to 35) and Fig. 6a shows that Co-RFSnooze achieves up to 43% savings (382^{nd} minute) compared to RFSnooze because higher number of user transfers allows switching off of additional RF chains (Section IIIB,C). Under low load conditions, Co-RFSnooze achieves lower savings of up to 29% (960^{th} minute) because (a) higher number of RF chains are switched off at individual

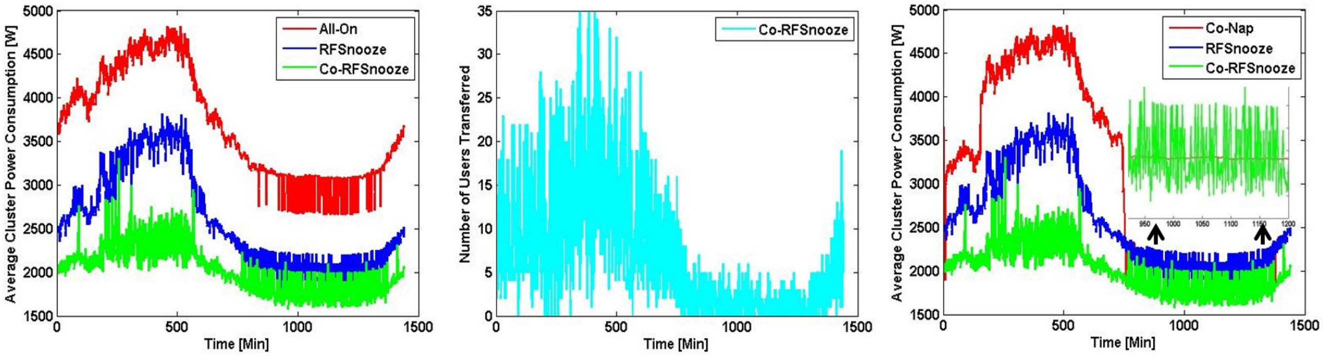

Fig. 6. (a) Comparison of average cluster power consumption of RFSnooze and Co-RFSnooze with that of All-On, (b) number of users transferred by Co-RFSnooze, and (c) comparison of average cluster power consumption of RFSnooze and Co-RFSnooze with that of Co-Nap

BSs by RFSnooze (b) the number of cluster users (Fig. 5b) and transferred users is lower as shown in Fig. 6b and (c) higher incidence of instances when no users are transferred resulting in identical performance of RFSnooze and Co-RFSnooze as indicated by corresponding instances in Fig. 6a.

Fig. 6c shows the P_C due to Co-Nap (shown in red), RFSnooze (shown in blue) and Co-RFSnooze (shown in green). Under high load, Co-Nap performance is comparable to All-On as it is unable to allow BSs to nap and satisfy the QoS constraints. RFSnooze achieves up to 35% gains (635^{th} minute) and Co-RFSnooze achieves up to 56% gains (382^{nd} minute) compared to Co-Nap under high load conditions. During transition from high load to low load and vice versa, Fig. 6c shows the dips in power consumption for Co-Nap (for instance between 50^{th} and 150^{th} minute) as lower load allows napping of BSs. RFSnooze and Co-RFSnooze outperform Co-Nap even in the transition regions by adapting BS resources and jointly adapting BS resources and UA respectively. The percentage of gains is lower compared to that under high load conditions at 22% (140^{th} minute) for RFSnooze and 38% (72^{nd} minute) for Co-RFSnooze. Under low load, Co-Nap outperforms RFSnooze as it is able to aggressively nap BSs and satisfy the QoS constraints. Co-RFSnooze outperforms Co-Nap whenever user transfers are possible which allows it to switch off additional RF chains. However, as explained earlier, whenever user transfers are not possible, Co-Nap outperforms Co-RFSnooze. The above behavior of Co-RFSnooze compared to Co-Nap is shown in the inset (zoomed-in section between 900^{th} and 1200^{th} minute) of Fig. 6c wherein the green curve repeatedly goes above and below the red curve. Also, due to the bulk of the savings coming from RFSnooze under low load, which underperforms Co-Nap, Co-RFSnooze achieves up to 11% (960^{th} minute) compared to Co-Nap.

Next, we will compare the number of cluster active RF chains used by the proposed algorithms with that used by All-On and Co-Nap in Figs. 7a and 7b respectively. The number of cluster active RF chains in (a) a frame is the sum of the active RF chains used at individual BSs and (b) a simulation step is the number of cluster active RF chains averaged over all the frames in the simulation step.

In Fig. 7a, all the cluster BS RF chains are active for All-On under high load whereas RFSnooze uses lesser number of RF chains and the least number are used by Co-RFSnooze. Under low load conditions, there are dips in the number of BS RF chains for All-On because there are no users associated with certain BSs in that instance and we see corresponding dips for RFSnooze and Co-RFSnooze as well. Fig. 7b shows that all the cluster RF chains are active for Co-Nap when the load is high as napping of BSs is not possible. Under low load, Co-Nap aggressively reduces the number of RF chains and thereby the power consumption as observed in Fig. 6c. RFSnooze consumes higher power than Co-Nap under low load conditions because it uses higher number of RF chains, as is evident from Fig. 7b. Further, we can see that the number of active RF chains used by Co-RFSnooze repeatedly goes above and below the number of RF chains used by Co-Nap. This results in similar pattern of P_C of Co-RFSnooze in Fig. 6c. During the transition from low load to high load and vice versa, the number of RF chains for RFSnooze and Co-RFSnooze is lower than that of Co-Nap. This is the cause for the trend of P_C of Co-Nap, RFSnooze and Co-RFSnooze during transition periods as seen in Fig. 6c.

Table VI presents the percentage of savings in P_C, averaged over 24 hours, for the proposed algorithms with respect to All-On and Co-Nap. Co-RFSnooze outperforms both All-On and Co-Nap when the savings are averaged over 24 hours which includes periods of low, medium and high loads.

We conclude the results by presenting the comparison of Co-RFSnooze and exhaustive search in Table VII. The simulation framework and parameters used is identical to

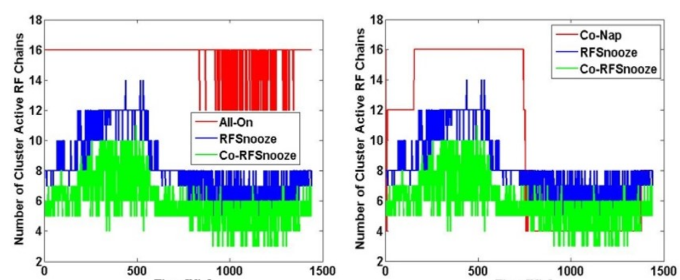

Fig. 7. Comparison of number of cluster active RF chains of RFSnooze and Co-RFSnooze with (a) All-On, and (b) Co-Nap

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2017.2738442, IEEE Transactions on Green Communications and Networking

TABLE VI
AVERAGE PERCENTAGE SAVINGS IN P_C OF RFSNOOZE AND CO-RFSNOOZE

	Low Load	High Load	Total
RFSnooze vs All-On	32.74%	26.21%	30%
Co-RFSnooze vs All-On	41.5%	47.38%	44.67%
RFSnooze vs Co-Nap	-16.1%	26%	7.68%
Co-RFSnooze vs Co-Nap	-0.86%	47.25%	25.52%

TABLE VII
AVERAGE PERCENTAGE SAVINGS IN P_C OF CO-RFSNOOZE COMPARED TO EXHAUSTIVE SEARCH

	Low Load	Medium Load	High Load
Co-RFSnooze vs Exhaustive Search	0%	-13%	-18%

that used for the remaining experiments except the following two changes. As the computational complexity of exhaustive search is exponential in $\mid I_C \mid$ (Section IIIE), to keep the simulation time tractable, we have chosen (a) the number of cluster users $\mid I_C \mid = 100$ and (b) low, medium and high load points of $0.1, 0.5, 0.8$ of the sample trace in Fig. 5a and the resulting number of users are $10, 50, 80$. We have conducted three runs of Co-RFSnooze and Exhaustive search for each of the load points and report the average percentage savings in P_C of Co-RFSnooze compared to exhaustive search in Table VII. The deviation of the Co-RFSnooze P_C from the optimal value achieved by exhaustive search is at most 18% at high load.

V. CONCLUSION

In this paper, we presented novel RF switching technique to minimize the average power consumption of a cluster of BSs in a transmission frame while satisfying the cluster users' QoS requirements and BS utilization constraints. Simulation results indicate that the proposed algorithms significantly outperform the conventional All-On scheme while Co-RFSnooze significantly gains over time slot based adaptive BS switching scheme Co-Nap under high and medium loads while being comparable under low load conditions.

REFERENCES

[1] R. Guruprasad, K. Son, and S. Dey, "Power-efficient base station operation through user QoS-aware adaptive RF chain switching technique," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 244–250.
[2] ERICSSON, "Ericsson mobility report on the pulse of networked society," Ericsson, Tech. Rep., November 2016.
[3] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 55–62, August 2011.
[4] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *CoRR*, vol. abs/1609.09773, 2016.
[5] NOKIA, "Flatten network energy consumption," Nokia Solutions and Networks, Tech. Rep., 2013.
[6] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modelling of base station power consumption under real traffic loads," *Sensors*, vol. 12, no. 4, pp. 4281–4310, 2012.
[7] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3292–3304, September 2012.
[8] Q. Wu, M. Tao, and W. Chen, "Joint Tx/Rx energy-efficient scheduling in multi-radio wireless networks: A divide-and-conquer approach," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2727–2740, April 2016.
[9] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4196–4209, August 2013.
[10] K. Son, S. Nagaraj, M. Sarkar, and S. Dey, "QoS-aware dynamic cell reconfiguration for energy conservation in cellular networks," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 2022–2027.
[11] K. Adachi, J. Joung, S. Sun, and P. H. Tan, "Adaptive coordinated napping (CoNap) for energy saving in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5656–5667, November 2013.
[12] Q. Zhang, C. Yang, H. Haas, and J. S. Thompson, "Energy efficient downlink cooperative transmission with BS and antenna switching off," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 5183–5195, September 2014.
[13] S. Han, C. Yang, and A. F. Molisch, "Spectrum and energy efficient cooperative base station doze," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 285–296, February 2014.
[14] D. W. K. Ng, Y. Wu, and R. Schober, "Power efficient resource allocation for full-duplex radio distributed antenna networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2896–2911, April 2016.
[15] X. Huang and N. Ansari, "Joint spectrum and power allocation for multi-node cooperative wireless systems," *IEEE Transactions on Mobile Computing*, vol. 14, pp. 2034–2044, 2015.
[16] Nokia, "3GPP setup of CoMP cooperation areas, R1-090725," NOKIA, Tech. Rep., February 2009.
[17] T. Biermann, "Dealing with backhaul network limitations in coordinated multi-point deployments," Ph.D. dissertation, Dept. Electrical Engineering, Paderborn University, Paderborn, Germany, 2012.
[18] A. Chatzipapas, S. Alouf, and V. Mancuso, "On the minimization of power consumption in base stations using on/off power amplifiers," in *2011 IEEE Online Conference on Green Communications*, September 2011, pp. 18–23.
[19] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas, "Minimizing base station power consumption," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 297–306, February 2014.
[20] ETSI, "Physical channels and modulation," TS 136 211, Tech. Rep. v.13.3.0, 2016.
[21] B. Han, J. Leblet, and G. Simon, "Hard multidimensional multiple choice knapsack problems, an empirical study," *Comput. Oper. Res.*, vol. 37, no. 1, pp. 172–181, Jan. 2010.
[22] ETSI, "LTE E-UTRA physical layer procedures," TS 36 213, Tech. Rep. v.13.0.0, 2016.
[23] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-QoS-aware fair scheduling for LTE," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.
[24] M. T. Kawser, N. I. B. Hamid, M. N. Hasan, M. S. Alam, and M. M. Rahman, "Downlink SNR to CQI mapping for different multiple antenna techniques in LTE," *International Journal of Information and Electronics Engineering*, vol. 2, no. 5, p. 757, 2012.
[25] A. Papadogiannis, E. Hardouin, and D. Gesbert, "A framework for decentralising multi-cell cooperative processing on the downlink," in *2008 IEEE Globecom Workshops*, November 2008, pp. 1–5.
[26] Y. Gao, Q. Wang, and G. Liu, "The access network and protocol design for CoMP technique in LTE-advanced system," in *2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)*, September 2010, pp. 1–4.
[27] K. Alexandris, N. Nikaein, R. Knopp, and C. Bonnet, "Analyzing X2 handover in LTE/LTE-A," in *2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2016, pp. 1–7.
[28] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: A practical interference management in heterogeneous wireless access networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1260–1272, June 2011.
[29] 3GPP, "Spatial channel model for multi input multi output (MIMO) simulations," 3GPP TR 25.996 2011-03, Tech. Rep. 25, 2011.
[30] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 56–61, June 2011.

Ranjini B Guruprasad Ranjini Guruprasad is currently a PhD student at the Department of Electrical and Computer Engineering, University of California, San Diego (UCSD). Her research interests lie in the design and analysis of algorithms for green networks, specifically with a focus on energy efficient operation of mobile devices and renewable energy powered base stations.

Sujit Dey Sujit Dey (SM'03âĂŞF'14) received the Ph.D. degree in computer science from Duke University, Durham, NC, USA, in 1991. He is a Professor in the Department of Electrical and Computer Engineering, University of California, San Diego (UCSD), La Jolla, CA, USA, where he heads the Mobile Systems Design Laboratory, which is developing innovative technologies in mobile cloud computing, adaptive multimedia and networking, green computing and communications, and predictive and prescriptive analytics to enable future applications in connected health, immersive multimedia, smart cities, and smart factories. He is the Director of the Center for Wireless Communications, and the Director of the Institute for the Global Entrepreneur at UCSD. He served as the Faculty Director of the von Liebig Entrepreneurism Center from 2013âĂŞ2015, and as the Chief Scientist, Mobile Networks, at Allot Communications from 2012âĂŞ2013. He founded Ortiva Wireless in 2004, where he served as its founding CEO and later as CTO and Chief Technologist until its acquisition by Allot Communications in 2012. Prior to Ortiva, he served as the Chair of the Advisory Board of Zyray Wireless till its acquisition by Broadcom in 2004, and as an advisor to multiple companies including ST Microelectronics and NEC. Prior to joining UCSD in 1997, he was a Senior Research Staff Member at NEC C&C Research Laboratories in Princeton, NJ, USA. He has co-authored more than 250 publications, and a book on low-power design. He holds 18 U.S. and two international patents, resulting in multiple technology licensing and commercialization. Dr. Dey has been a recipient of six IEEE/ACM Best Paper Awards, and has chaired multiple IEEE conferences and workshops.