

# Adaptive Coordinated Napping (CoNap) for Energy Saving in Wireless Networks

Koichi Adachi, Jingon Joung, Sumei Sun, and Peng Hui Tan

**Abstract**—We propose a time slot based transmission strategy, referred to as *adaptive coordinated napping (CoNap)*, for energy saving in cellular networks under time-varying traffic demand. In adaptive CoNap network, multiple neighboring base stations (BSs) form a cluster and each BS operates in either a transmit mode (TM) or a nap mode (NM) in each time slot. The dynamic assignment of TM and NM to each BS is implicitly coordinated among multiple BSs. This implicit coordination is realized by a binary general flickering pattern matrix (FPM) through adaptively selected mapping matrix (MM). To track the time-varying traffic demand, we develop an adaptive algorithm to dynamically select the appropriate MM from a predefined MM set by taking into account the network quality of service (QoS) requirement. Our numerical results based on a realistic energy consumption model in a cellular network show that as high as 40% saving can be achieved without compromising the network QoS.

**Index Terms**—Cellular networks, coordinated base station napping, dynamic base station switching, energy saving, green communications, resource scheduling, sleeping strategy.

## I. INTRODUCTION

ENERGY efficiency (EE) has become important performance metric of cellular communication systems due to the high demand for energy saving [1]. The optimization objective in the communication system design has been shifted from high spectrum efficiency (SE) to high EE [1]–[3], and a good tradeoff between SE and EE [4].

The traffic demand of cellular systems has a dynamic nature in both time and space [5]. Under a *high traffic demand condition*, most of the radio resources such as time slots, frequency bands, and multiple antennas are utilized with a high probability. On the other hand, under a *low traffic demand condition*, only part of the resources is necessary to satisfy the traffic demand. Thus, there is a large potential for energy saving. The base station (BS) is a major power consumer of the whole network [6], hence network energy saving can be effectively achieved by reducing BS's power consumption. This observation leads us to consider a *BS deactivation* method.

Manuscript received November 27, 2012; revised March 27 and July 23, 2013; accepted August 16, 2013. The associate editor coordinating the review of this paper and approving it for publication was S. Affes.

The authors are with the Institute for Infocomm Research, A\*STAR, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632 (e-mail: {kadachi, jgjong, sunsm, phtan}@i2r.a-star.edu.sg).

Part of this work has been published in the Proceedings of the IEEE Global Communications Conference (GLOBECOM) Workshop, Anaheim, USA, Dec. 2012.

Digital Object Identifier 10.1109/TWC.2013.101613.121873

## A. Related Works

The BS activation and deactivation can be adapted either in a long time period (e.g., hour-scale) which is called a *long duration approach* (LDA) [7]–[13], or in a short time period (e.g., millisecond-scale) which is called a *short duration approach* (SDA) [14]–[16] as shown in Table I.

1) *Long Duration Approach (LDA)*: In LDA, a BS is deactivated (or switched to a sleep mode) for a long period of time. The coverage area of the deactivated BS is called *dormant coverage area*. The users within the dormant coverage area are serviced by nearby active BSs to guarantee their quality-of-service (QoS) requirement. Hence, two main issues need to be addressed: how we select appropriate BSs to be deactivated and when we deactivate/activate BSs.

In [8], the traffic load<sup>1</sup> is used as a metric to select the BS to be deactivated. The less the traffic load is, the less the BS is utilized. The underutilized BS is deactivated by handing over its traffic to the neighboring BSs. Both centralized/decentralized algorithms are proposed for the selection [8]. Once BS is deactivated, the neighboring active BSs need to support the dormant coverage area. To guarantee QoS requirement, a cell zooming method [8] and a coordinated multipoint (CoMP) method [10] have been mainly considered. In cell zooming, active BSs expand their coverage. In CoMP, the dormant coverage area is covered by multiple active BSs via coordinated beamforming [17], [18]. A simple power control (PC) algorithm combined with cell zooming/CoMP is proposed to increase the EE in [12]. In [13], the distance information between the BS and users is used to select the BSs to be deactivated. The BS with the longest average distance to users is deactivated. In cell zooming, the active BS expands its coverage by either increasing transmit power or optimizing vertical antenna tilt angle in order to support the dormant coverage area. If the former approach is adopted, the power amplifier (PA) needs to be carefully designed as the PA efficiency (PAE) is dependent on the output power [19]. Generally, a PA is designed so that the maximum PAE is obtainable at the maximum output power. To increase the transmit power during cell zooming period, the PA characteristic should be matched to that period<sup>2</sup>. This results in poor PAE during non-cell zooming period. If antenna tilt angle adaptation is used, the change of antenna tilt angle

<sup>1</sup>In this paper, we explicitly differentiate two terminologies: *traffic demand* (or simply *traffic*) and *traffic load*. Traffic demand indicates the amount of traffic arriving into the system, e.g., the number of users. The traffic load indicates the actual resource utilization. The traffic load is calculated as a ratio of the system resource (time, frequency etc) currently occupied to the available total system resource.

<sup>2</sup>Power amplifier switching (PAS) configuration may be a solution [20].

may introduce large inter-cell interference (ICI) to neighboring coverage area supported by other active BSs. Channel state information (CSI) and user data exchanges are essential for CoMP, hence high capacity backhaul link is required [21]. In addition, the complex optimization for beamforming is necessary [10].

In [22], the *potential energy saving gain* obtained by optimizing the BS density according to the user density is shown. The time-varying nature of the traffic demand is used to decide when to deactivate and reactivate the BSs [7], [9]. For simplicity, let us consider the homogeneous network condition, i.e., the traffic demand at each BS is equal. The traffic at time instance  $t$  normalized by its peak value is denoted by  $0 \leq f(t) \leq 1$ , which is modeled as a sinusoidal wave. Suppose a fraction  $x < 1$  of BSs within the system remain active, while the remaining  $(1 - x)$  BSs are deactivated if the following is satisfied [7], [9]:

$$f(t) \cdot \left(1 + \frac{1-x}{x}\right) = \frac{1}{x}f(t) < f_{th}, \quad (1)$$

where  $f_{th}$  is a predetermined threshold. To balance the QoS degradation and energy saving,  $f_{th}$  needs to be carefully set [9]. The deactivated BS is to be reactivated when the approximated traffic becomes  $f(t) > xf_{th}$  [7] or  $f(t) > f_{th}$  [9]. In (1), it is assumed that the same amount of traffic  $\frac{1-x}{x}f(t)$  is handed over to each remaining active BS. For a given traffic demand, however, the actual traffic load may vary significantly due to the surrounding environment such as channel condition. The assumption made in (1) does not take into account this important issue.

The existing LDAs have several drawbacks: i) insufficient adaptation to the dynamic nature of traffic demand due to the long operation cycle, ii) a number of parameter optimizations such as antenna tilt angle and traffic demand threshold value, iii) huge overhead and complex operation for cooperation among BSs, and iv) reduced PAE due to the increase of transmit power during cell zooming period.

2) *Short Duration Approach (SDA)*: The PA is major power consumer in BS [23]. When traffic load is zero, it is possible to put PA into micro sleep mode which requires significantly low power [2]. In [2], a simple on-off scheduler is proposed to increase EE. The scheduler works to let the subframes empty as often as possible to facilitate micro sleep. In [14], cell discontinuous transmission (DTX) is proposed to reduce power consumption of 3rd generation partnership project (3GPP) long-term evolution (LTE) system. Cell DTX makes use of multicast and broadcast single frequency network (MBSFN) subframe if traffic demand is low. In 3GPP LTE, time domain inter-cell interference coordination (ICIC) is introduced for heterogeneous network (HetNet) to reduce the interference from macro BS to users in a small cell by muting certain subframes [24]. This interference reduction can contribute to the power saving. However, the configuration of MBSFN subframes in cell DTX is performed by each BS, i.e., no coordination among BSs. Thus, cell DTX does not fully facilitate the interference reduction, which can be obtained by coordination among BSs.

In [15], it is shown that the coordination of sleep mode among neighboring BSs can bring EE improvement due to the ICI reduction under *fixed traffic demand condition*. Although

coordination is considered in [15], the number of coordinated BSs is restricted to three and no adaptive selection of coordinated sleep pattern is considered. Furthermore, the time-varying traffic demand is not considered. In [16], a traffic demand-aware PC combined with BS sleep is proposed to maximize EE. The authors propose two strategies to determine when to activate the deactivated BS. The number of awaiting users in the queue and predetermined vacation duration are used as the criteria. To the best of authors' knowledge, no literature has considered adaptive coordination among multiple BSs under time-varying traffic demand for SDA.

## B. Proposed Approach and Contributions

To address the time-varying dynamic traffic demand and solve the problem of existing BS deactivation approaches, we propose a time slot based transmission strategy which is termed by *adaptive coordinated napping (CoNap)*. Adaptive CoNap belongs to SDA. The napping is equivalent to the subframe-based sleeping in [15], yet it represents well the behavior of BS which transmits after very short non-transmission time (e.g., 1ms for subframe of LTE system). In adaptive CoNap networks, multiple neighboring BSs form a cluster. Each BS in a cluster operates in either a transmit mode (TM) or a nap mode (NM) at each time slot. In TM, a BS is allowed to transmit the signals to its associated users by allocating frequency domain RBs, and the actual transmission is determined by a resource allocation algorithm. The BS may not transmit any signal to users even during TM time slot. In this case, transmission overhead arises partially as the time slot is declared TM time slot. In NM, a BS is predetermined not to transmit any signal, and stands by for the quick transition to TM. The neighboring BSs in NM do not generate interference, hence the overall network interference is reduced. The operation of each BS is represented by a binary vector with a value of "1" for TM and "0" for NM, which is called *flickering pattern*. The flickering patterns for BSs in a cluster can be represented by a general flickering pattern matrix (FPM) and a mapping matrix (MM).

The proposed adaptive CoNap supports arbitrary number of coordinated BSs and flexible flickering pattern assignment by the use of the general FPM and the MM. In this paper, we consider two types of QoS: user QoS, i.e., target rate, and network QoS which is the probability that users have their QoS requirement satisfied. We need to dynamically select the flickering pattern to achieve energy saving without compromising the user and network QoS under time-varying traffic demand. This is realized by selecting the appropriate MM from the set of available MMs based on the current traffic load. This enables the energy efficient multi-cell (BS) coordination. As a result, the number of users whose QoS (target rate) can not be satisfied is kept below a certain number, while achieving energy saving.

Our proposed adaptive CoNap has the following advantages over the existing LDAs and SDAs: i) fine tracking ability of the dynamic traffic demand, ii) no complicated cooperation among BSs, but still implicit coordination, iii) no transmit power increase nor antenna tilt angle adaptation, iv) no coverage expansion for BSs, v) no additional control signaling between BS and users.

TABLE I  
RELATIONSHIP BETWEEN PREVIOUS WORKS AND THIS WORK

|                      | Long duration approach (LDA) |                                                                                                                                 | Short duration approach                                                   |                                 |
|----------------------|------------------------------|---------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------|---------------------------------|
|                      | w/o coordination             | w/ coordination                                                                                                                 | w/o coordination                                                          | w/ coordination                 |
| Fixed Traffic Demand | –                            | BS sleeping w/ CoMP [10]                                                                                                        | –                                                                         | Coordinated sleep schedule [15] |
| Dynamic traffic      | –                            | Cell zooming [7], [8], [11]<br>Dynamic BS switching [9]<br>Inter-cell cooperation w/ PC [12]<br>Distance-aware BS sleeping [13] | On-off scheduler [2]<br>Cell DTX [14]<br>Load-aware PC with BS sleep [16] | Adaptive CoNap                  |

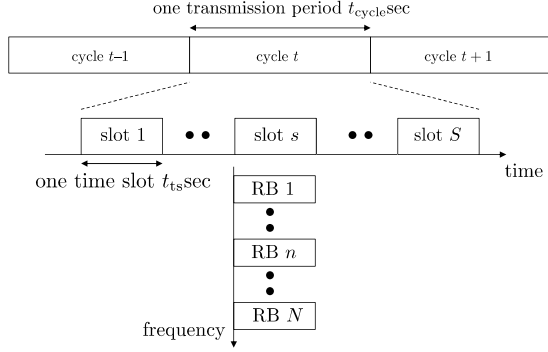


Fig. 1. Frame structure consisting of  $S$  orthogonal time slots and  $N$  frequency domain RBs.

To properly verify the overall energy saving gain achieved by adaptive CoNap, we propose a practical BS power consumption model by taking into account the traffic load and the imperfect PAE. The simulation results show that as high as 40% energy saving can be achieved by adaptive CoNap with negligible QoS degradation.

The rest of the paper is organized as follows. In Section II, the proposed system model and BS power consumption model are given. After describing the energy saving via the proposed CoNap in Section III, the adaptive algorithm for MM selection is proposed in Section IV. The simulation results are presented in Section V. Finally, Section VI concludes the paper.

## II. PROPOSED SYSTEM MODEL

We consider a downlink orthogonal frequency division multiplexing (OFDM) based cellular system. The whole frequency bandwidth  $W_{\text{total}}$  Hz is divided into multiple RBs with equal bandwidth as shown in Fig. 1. The total number of frequency domain RBs is denoted by  $N$ , i.e., the frequency bandwidth of each RB is  $W_{\text{RB}} \triangleq \frac{W_{\text{total}}}{N}$  Hz. The total number of time slots in one transmission cycle is denoted by  $S$ . The durations of one transmission cycle and one time slot are  $t_{\text{cycle}}$  and  $t_{\text{ts}} (\triangleq \frac{t_{\text{cycle}}}{S})$ , respectively. For example, one time slot can be set to 1ms, the minimum transmit time interval (TTI) of LTE system. An *admission control* selects the users to be served during each transmission cycle based on some predetermined criterion such as target blocking rate. The selected user set within the cluster during transmission cycle  $t$  is denoted by  $\mathcal{U}(t)$ .

In the proposed adaptive CoNap network, each BS operates in either TM or NM at each time slot according to the flickering pattern, which will be explained later. If each BS randomly selects its flickering pattern, the ICI may not be reduced effectively. We are therefore motivated to consider grouping the BSs into the clusters. The BSs in the same cluster

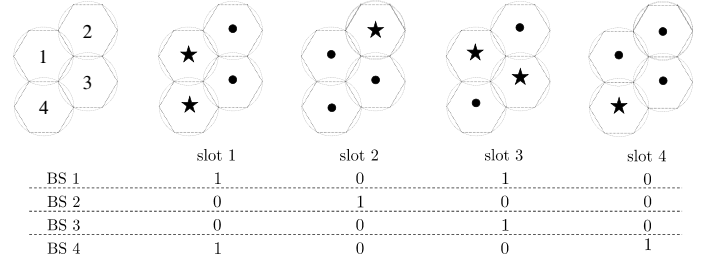


Fig. 2. An example of BS flickering over three time slots. ‘★’ and ‘●’ represent transmit mode (TM) and nap mode (NM), respectively.

share the flickering pattern assignment information.  $|\mathcal{B}_{\text{total}}|$  BSs are divided into non-overlapping clusters with each cluster consisting of multiple BSs, where  $|\mathcal{X}|$  denotes the cardinality of  $\mathcal{X}$ . Without loss of generality, each cluster is assumed to consist of  $B$  BSs. We refer to  $B$  as a cluster size.

In the following, let us focus on one cluster and denote the BS set within that cluster by  $\mathcal{B}$  and other BS sets by  $\bar{\mathcal{B}}$ , i.e.,  $|\mathcal{B}| = B$ ,  $\mathcal{B} \cup \bar{\mathcal{B}} = \mathcal{B}_{\text{total}}$ , and  $\mathcal{B} \cap \bar{\mathcal{B}} = \emptyset$ . The number of time slots during one transmission cycle is also the flickering pattern cycle. In Fig. 2, CoNap is illustrated with a cluster size of four and the flickering pattern cycle of four time slots, i.e.,  $B = 4$  and  $S = 4$ . The first BS, BS1, is in TM mode at time slot 1 and 3, and in NM mode at time slot 2 and 4, i.e., the flickering pattern for BS1 is (1 0 1 0).

### A. Binary General Flickering Pattern Matrix (FPM) and Mapping Matrix (MM)

A binary general FPM is a  $Q$ -by- $S$  binary matrix and denoted by  $\mathbf{F}_G = (\mathbf{f}_1^T \cdots \mathbf{f}_q^T \cdots \mathbf{f}_Q^T)^T$  where  $(\cdot)^T$  denotes the transpose operation.  $\mathbf{F}_G$  captures the operation of BSs during flickering pattern cycle  $S$ . Since the element of general FPM  $\mathbf{F}_G$  is binary (either “1” or “0”), the largest value of  $Q$  is  $2^S$ .  $\mathbf{f}_q = (f_{q,1} \cdots f_{q,s} \cdots f_{q,S})$ ,  $q \in \{1, \dots, Q\}$ , which is the  $q$ th row of  $\mathbf{F}_G$ , is the  $q$ th flickering pattern of length  $S$ . The element  $f_{q,s} \in \{1, 0\}$  denotes the BS operation mode at time slot  $s$ : if  $f_{q,s} = 1$ , the BS is in TM, and otherwise, it is in NM. Note that  $\mathbf{F}_G$  is not an orthogonal matrix, and for each  $\mathbf{f}_q$ , only a subset of the  $Q$  vectors are orthogonal to it.

Let us denote the FPM for the cluster as  $\mathbf{J} = (\mathbf{j}_1^T \cdots \mathbf{j}_b^T \cdots \mathbf{j}_B^T)^T$  where  $\mathbf{j}_b$  is a length  $S$  binary row vector representing the flickering pattern assigned to BS  $b \in \mathcal{B}$ . A  $B$ -by- $Q$  binary MM  $\mathbf{M}_m \in \mathcal{M}$ , where  $\mathcal{M}$  is the set of available MMs, is defined as

$$\mathbf{M}_m \triangleq (\mathbf{e}_{i_1(m)}^T \cdots \mathbf{e}_{i_b(m)}^T \cdots \mathbf{e}_{i_B(m)}^T)^T, \quad (2)$$

where  $\mathbf{e}_{i_b(m)}$  is a length  $Q$  row vector with its  $i$ th element being “1” and the rest elements being “0”. Hence when the



$j$ th element of  $b$ th row vector is “1”, the  $j$ th flickering pattern vector is assigned to BS  $b$ .

Using the general FPM and the MM in (2), we can determine the FPM as follows:

$$\mathbf{J} = \underbrace{(\mathbf{e}_{i_1(m)}^T \cdots \mathbf{e}_{i_b(m)}^T \cdots \mathbf{e}_{i_B(m)}^T)}_{\mathbf{M}_m} \underbrace{(\mathbf{f}_1^T \cdots \mathbf{f}_q^T \cdots \mathbf{f}_Q^T)}_{\mathbf{F}_G}^T = (\mathbf{f}_{i_1(m)}^T \cdots \mathbf{f}_{i_b(m)}^T \cdots \mathbf{f}_{i_B(m)}^T)^T. \quad (3)$$

The flickering pattern  $\mathbf{f}_{i_b(m)}$  is assigned to BS  $b$ . An MM,  $\mathbf{M}_m$ , is selected from the set of MMs  $\mathbf{M}_m \in \mathcal{M}$  to realize a certain flickering pattern in a cluster.

### B. Received SINR at Users

Let us suppose the system is operating with MM  $\mathbf{M}_m \in \mathcal{M}$  during transmission cycle  $t$ . The average received SINR at user  $u \in \mathcal{U}$  from BS  $b$  during time slot  $s$  in transmission cycle  $t$  is calculated as

$$\gamma_{u,b}(t, m, s) = \frac{f_{i_b(m),s} P_{\text{TX}} \Gamma_{u,b}(t)}{I_{u,b}(t, m, s) + \sigma_n^2}, \quad (4)$$

where  $f_{i_b(m),s} \in \{1, 0\}$  indicates that BS  $b$  is in TM if  $f_{i_b(m),s} = 1$  and in NM if  $f_{i_b(m),s} = 0$ ,  $P_{\text{TX}}$  is the transmit power per unit bandwidth,  $\Gamma_{u,b}(t)$  is the channel gain between user  $u$  and BS  $b$  which is assumed to be constant during one cycle,  $I_{u,b}(t, m, s)$  is the interference from neighboring BSs, and  $\sigma_n^2$  is the additive white Gaussian noise (AWGN) power. In the following, we omit the transmission cycle index  $t$  unless otherwise required.

In this paper, only the large scale fading effects, i.e., distance dependent path-loss and shadowing-loss, are considered. The channel gain,  $\Gamma_{u,b}$ , is given by

$$\Gamma_{u,b} = 10^{\frac{A(\psi_{u,b}) + L_{u,b} + \eta_{u,b}}{10}}, \quad (5)$$

where  $A(\psi_{u,b})$  is the antenna gain with  $\psi_{u,b}$  being the elevation angle of user  $u$  from the main-beam direction of the antenna of BS  $b$ ,  $L_{u,b}$  is the distance dependent path-loss, and  $\eta_{u,b}$  is the log-normally distributed shadowing-loss between user  $u$  and BS  $b$  with the standard deviation of  $\sigma$  (dB).

The interference  $I_{u,b}(m, s)$  is defined as

$$I_{u,b}(m, s) \triangleq P_{\text{TX}} \sum_{b' \in (\mathcal{B}_{\text{total}} \setminus b)} f_{i_{b'},s}(m) \Gamma_{u,b'}. \quad (6)$$

### C. Frequency Domain RB Allocation and Traffic Load of Each BS

We consider the target rate as a user QoS requirement. Since there is no explicit cooperation among BSs, the RB allocation information at each BS is not shared among the BSs. Hence, each BS allocates sufficient number of RBs to its users to achieve the target rate under the assumption of the largest interference. In other word, the BS may over allocate the RBs to its users.

We define the target rate of user  $u$  as  $R_u^{\text{tar}} \triangleq \frac{V_u}{t_{\text{cycle}}}$ , where  $V_u$  denotes the number of bits need to be delivered during one transmission cycle. Let  $n_{u,b}(m, s) \in \{0, 1, \dots, N\}$  indicate the number of RBs at BS  $b$  allocated to user  $u$  during time

slot  $s$ . To successfully deliver  $V_u$  bits during one transmission cycle, we need to satisfy

$$t_{\text{cycle}} R_u^{\text{tar}} \leq t_{\text{ts}} \sum_{b \in \mathcal{B}} \sum_{s=1}^S n_{u,b}(m, s) R_{u,b}(m, s) \quad (\text{bits}). \quad (7)$$

Accordingly, we obtain

$$R_u^{\text{tar}} \leq \frac{1}{S} \sum_{b \in \mathcal{B}} \sum_{s=1}^S n_{u,b}(m, s) R_{u,b}(m, s) \quad (\text{bps}), \quad (8)$$

where  $R_{u,b}(m, s)$  is the achievable rate of user  $u$  per RB by connecting to BS  $b$  during time slot  $s$ , which is calculated as

$$R_{u,b}(m, s) = W_{\text{RB}} \log_2(1 + \gamma_{u,b}(m, s)) \quad (\text{bps}). \quad (9)$$

Due to the limited number of RBs, some of the users may not have (8) satisfied. Those users are considered to be *blocked*. The blocking rate of the system during one transmission period with  $\mathbf{M}_m$  is defined as  $P_{\text{blk}}(m) \triangleq \frac{|\mathcal{U}_{\text{blk}}(m)|}{|\mathcal{U}|}$ , and  $\mathcal{U}_{\text{blk}}(m)$  is the set of blocked users.

The traffic load  $\rho_b(m, s)$  is defined as the fraction of the RBs occupied during time slot  $s$ , which is calculated as

$$\rho_b(m, s) = \frac{1}{N} \sum_{u \in \mathcal{U}} n_{u,b}(m, s). \quad (10)$$

### D. BS Energy Consumption Model

About 50 – 80% of power at BS is consumed at PA [23]. Thus, the overall power consumption depends on PAE. The typical PAE is assumed to be around 38% [2], however it depends on the output transmit power [19]. The transmit power depends on the traffic load. We therefore adopt a practical power consumption model to capture the contribution of both traffic load and imperfect PAE. Doherty PAs provide a high PAE and are widely employed in BSs [2], hence we focus on Doherty PA in this paper.

The radio frequency (RF) transmit power is expressed as a function of  $\rho_b(m, s)$  as [2]

$$\begin{aligned} P_{\text{RF}}(\rho_b(m, s)) &= \underbrace{(P_{\text{max}} - P_{\text{OH}})}_{=W_{\text{total}} P_{\text{TX}}} \rho_b(m, s) + P_{\text{OH}} \kappa(f_{i_b(m),s}, \rho_b(m, s)), \end{aligned} \quad (11)$$

where  $P_{\text{max}}$  is the maximum transmit power and  $P_{\text{OH}} = p_{\text{OH}} P_{\text{max}}$  ( $0 \leq p_{\text{OH}} < 1$ ) is the fixed overhead of the radiated power related to essential signals for the system, e.g., a reference signal and a control signal, and

$$\kappa(x, y) = \begin{cases} 0 & x = 0, \\ 0.5 & x = 1 \text{ and } y = 0, \\ 1 & x = 1 \text{ and } y \neq 0. \end{cases} \quad \begin{aligned} (12a) \\ (12b) \\ (12c) \end{aligned}$$

(12a) is for NM. Since BS is in NM, no overhead is necessary. (12b) represents the situation when BS is in TM but no user is associated. Since the BS is in TM, overhead should be generated partially. (12c) is for the situation when BS is in TM and there are users associated with it.

The PAE is defined as the ratio of output RF power to required DC power. The Doherty PAE is expressed as [19] (The derivation is given in Appendix)

$$\eta(P_{\text{RF}}(\rho_b(m, s))) = \begin{cases} \frac{\pi}{2} \sqrt{cP_{\text{RF}}(\rho_b(m, s))} & \text{for } 0 < cP_{\text{RF}}(\rho_b(m, s)) \leq \frac{1}{4}, (13a) \\ \frac{\pi}{2} \frac{cP_{\text{RF}}(\rho_b(m, s))}{3\sqrt{cP_{\text{RF}}(\rho_b(m, s))} - 1} & \text{for } \frac{1}{4} < cP_{\text{RF}}(\rho_b(m, s)) \leq 1, (13b) \end{cases}$$

where  $c = (\xi P_{\text{max}})^{-1}$  with  $\xi$  being the PA output backoff (OBO). The OBO is determined by the peak-to-average power ratio (PAPR) in order to avoid the nonlinear region of PA.

By denoting the amplification gain of the PA by  $g$ , the input signal power to PA is given by  $P_{\text{in}}(\rho_b(m, s)) = P_{\text{RF}}(\rho_b(m, s))/g$ . Then, the total power consumed at PA for output power of  $P_{\text{RF}}(\rho_b(m, s))$  is expressed as

$$P_{\text{PA}}(\rho_b(m, s)) = \left( \frac{1}{g} + \frac{1}{\eta(P_{\text{RF}}(\rho_b(m, s)))} \right) P_{\text{RF}}(\rho_b(m, s)), \quad (14)$$

and the BS power consumption for given traffic load  $\rho_b(m, s)$  is

$$P(\rho_b(m, s)) = (P_{\text{fix}} + \rho_b(m, s)P_{\text{dyn}} + P_{\text{PA}}(\rho_b(m, s)))p_{\text{loss}}, \quad (15)$$

where  $P_{\text{fix}}$  and  $P_{\text{dyn}}$  denote the power consumption at small-signal RF transmitter excluding the PA and base band interface, respectively, and  $p_{\text{loss}}$  is the total power loss incurred by an AC-DC unit for connection to electrical power grid, a DC-DC power supply for providing stable DC power to each unit, and an active cooling system [25]. The power consumption during NM is given as

$$P_{\text{NM}} \triangleq p_{\text{loss}}P_{\text{fix}}. \quad (16)$$

Eventually, the total energy consumption of a cluster during one cycle with MM  $\mathbf{M}_m$  is given as

$$E_c(m) = t_{\text{ts}} \sum_{b \in \mathcal{B}} \sum_{s=1}^S P(\rho_b(m, s)). \quad (17)$$

### III. RB ALLOCATION

In this section, we propose a network energy saving RB allocation for the CoNap network.

#### A. Frequency Domain RB Allocation

To minimize the total energy consumption per cluster, the optimization problem for a given MM  $\mathbf{M}_m$  is formulated as

$$\min_{n_{u,b}(m,s)} E_c(m) \quad (18a)$$

$$\text{s.t.} \quad \frac{1}{S} \sum_{b \in \mathcal{B}} \sum_{s=1}^S n_{u,b}(m, s) R_{u,b}(m, s) \geq R_u^{\text{tar}}, \quad \forall u, \quad (18b)$$

$$\sum_{u \in \mathcal{U}} n_{u,b}(m, s) \leq N, \quad \forall b, s, \quad (18c)$$

$$\sum_{b \in \mathcal{B}} \nu \left( \sum_{s=1}^S n_{u,b}(m, s) \right) \leq 1, \quad \forall u. \quad (18d)$$

The function  $\nu(x)$  is defined as

$$\nu(x) \triangleq \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \quad (19)$$

Constraint (18b) guarantees each user's target rate in each transmission cycle, (18c) guarantees that the number of allocated RBs during each time slot at each BS is no larger than the total number of RBs, and (18d) follows the fact that each user is associated with at most one BS at a time.

Problem (18) is a combinatorial problem with the complexity  $O(B^{|\mathcal{U}|})$ , we develop a suboptimum algorithm to solve it. Note that any resource allocation algorithm can be incorporated into the proposed adaptive CoNap network. To this end, the following observation can be used. While satisfying (18b),  $n_{u,b}(m, s)$  is minimized by connecting user  $u$  to the BS with the highest  $\gamma_{u,b}(m, s)$  as it is a decreasing function over  $\gamma_{u,b}(m, s)$ . Then, the following suboptimum algorithm is developed to solve the optimization problem (18).

- Step 1: Associate each user with the BS which has the highest received SINR, i.e.,  $b_u^* = \arg \max_{b \in \mathcal{B}} (\max_s \gamma_{u,b}(m, s))$ . The set of users associated with BS  $b$  is denoted by  $\mathcal{U}_b$ . The total rate of each user is initialized to  $R_u = 0$  for  $u \in \mathcal{U}_b$ . Repeat step 2 – 4 for all the TM time slots.
- Step 2: Initialize the number of remaining RBs of BS  $b$  at time slot  $s$  as  $N_b(s) = N$ .
- Step 3: Select user  $u^* = \arg \max_{u \in \mathcal{U}_b} \gamma_{u,b}(m, s)$  and calculate the required number of RBs as

$$n_{u^*,b}(m, s) = \left\lceil \frac{(R_{u^*}^{\text{tar}} - R_{u^*})S}{R_{u^*,b}(m, s)} \right\rceil, \quad (20)$$

where  $\lceil x \rceil$  denotes the smallest integer greater than or equal to  $x$ .

- Step 4: If  $n_{u^*,b}(m, s) \leq N_b(s)$ , allocate  $n_{u^*,b}(m, s)$  RBs to user  $u^*$  and decrease  $N_b(s)$  by  $n_{u^*,b}(m, s)$ . Update the user set as  $\mathcal{U}_b = (\mathcal{U}_b \setminus u^*)$  and go back to Step 3. Otherwise, allocate  $N_b(s)$  RBs to user  $u^*$ , update the total rate as  $R_{u^*} = R_{u^*} + N_b(s)R_{u^*,b}(m, s)/S$  and set  $N_b(s) = 0$ . Move to the next allocated time slot and go back to step 2.
- Step 5: If  $R_u < R_u^{\text{tar}}$ , consider user  $u$  as blocked, i.e.,  $\mathcal{U}_{\text{blk}}(m) = (\mathcal{U}_{\text{blk}}(m) \cup u)$ .

#### B. Flickering Pattern Mapping Matrix (FPMM)

1) *General-Flickering Pattern Mapping Matrix (G-FPMM)*: We consider *general-flickering pattern mapping matrix (G-FPMM)*. For G-FPMM, consecutive time slots are allocated to each BS in order to reduce the energy consumption of user [27]. The G-FPMM is a compromised solution between the ICI reduction and the number of available time slots for each BS.

G-FPMM allocates  $m$  consecutive TM time slots to each BS, hence the flickering patterns of BSs may not be orthogonal to each other. Since the total number of time slots in one transmission cycle is limited to  $S$ , consecutive time slots may only be allocated to some BSs in consecutive transmission cycles. Then, the position of the  $s$ th TM time slot for BS  $b$

is uniquely expressed as  $(b + s - 1) - \lfloor \frac{b+s-1}{S+1} \rfloor S$ . Thus, the following G-FPMM is obtained:

$$\mathbf{M}_m = (\mathbf{e}_{i_1(m)}^T \cdots \mathbf{e}_{i_b(m)}^T \cdots \mathbf{e}_{i_B(m)}^T)^T \quad (21)$$

with

$$i_b(m) = 1 + \sum_{s=1}^m 2^{S-\{(b+s-1)-\lfloor \frac{b+s-1}{S+1} \rfloor S\}}. \quad (22)$$

For example, for  $S = B = 4$  and  $m = 2$ , we have

$$\mathbf{J}_2 = \underbrace{(\mathbf{e}_{13}^T \mathbf{e}_7^T \mathbf{e}_4^T \mathbf{e}_{10}^T)^T}_{\mathbf{M}_2} \mathbf{F}_G = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}. \quad (23)$$

2) *Orthogonal-Flickering Pattern Mapping Matrix (O-FPMM)*: One special case of G-FPMM is *orthogonal-flickering pattern mapping matrix (O-FPMM)* which is obtained by setting  $m = 1$  in (22). From Remarks 1 and 2 in [26], the O-FPMM can completely eliminate the intra-cluster interferences and reduce the inter-cluster interferences, resulting in energy saving of the system. However, the number of available time slots for each BS is reduced by  $S$ , this may bring the increase of blocking rate.

The O-FPMM needs to satisfy the orthogonality among the BSs within a cluster and which is expressed as

$$\mathbf{M}_1 = (\mathbf{e}_{i_1(1)}^T \cdots \mathbf{e}_{i_b(1)}^T \cdots \mathbf{e}_{i_B(1)}^T)^T \quad (24)$$

with

$$i_b(1) = 1 + 2^{S-b}. \quad (25)$$

For example, when  $S = B = 4$ , we have the following FPM:

$$\mathbf{J}_1 = \underbrace{(\mathbf{e}_9^T \mathbf{e}_5^T \mathbf{e}_3^T \mathbf{e}_2^T)^T}_{\mathbf{M}_1} \mathbf{F}_G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (26)$$

3) *Non-Flickering Pattern Mapping Matrix (N-FPMM)*: Another special case of G-FPMM is *non-flickering pattern mapping matrix (N-FPMM)*, i.e., there is no coordination among BSs. The N-FPMM  $\mathbf{M}_S$  is obtained by setting  $m = S$  in (22) and we have  $i_b(S) = 1 + 2^S$ . By setting  $S = B$ , we have

$$\mathbf{J}_B = \mathbf{1}_B, \quad (27)$$

where  $\mathbf{1}_B$  denotes a  $B$ -by- $B$  matrix whose all elements are ones. By solving the aforementioned optimization problem with  $\mathbf{J}_B$ , we have a similar solution to on-off scheduling in [2]. For N-FPMM, all the BSs are in TM for every time slot and each individual BS performs resource allocation independently. Since the scheduler tries to squeeze the user's signal to a smaller time slot index, neighboring BSs transmit the signals in the same time slots with a high probability. Thus, the ICI can be severe as there is no coordination among BSs. In the worst case, all the BSs transmit the signals to their associated users simultaneously, i.e., in the same time slot(s).

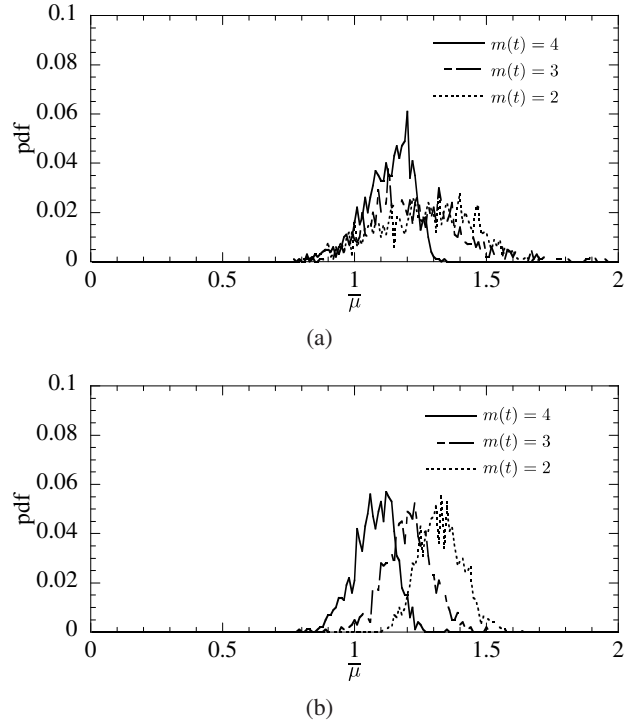


Fig. 3. The probability density function (pdf) of  $\bar{\mu}$  for  $B = 4$  (a)  $|\mathcal{U}|/B = 10$  (b)  $|\mathcal{U}|/B = 20$ .

#### IV. ADAPTIVE FLICKERING PATTERN SELECTION

The flickering pattern needs to adapt to the time-varying traffic demand. This can be realized by an adaptive MM selection algorithm among G-FPMM  $\mathbf{M}_m$  with  $m \in \{1, \dots, B\}$ . We denote the set of available MMs by  $\mathcal{M} = \{\mathbf{M}_1, \dots, \mathbf{M}_B\}$ .

Two conditions need to be satisfied during MM selection:

- C1:** No violation of network QoS (blocking rate) constraint.
- C2:** Network energy saving gain.

The optimum selection requires the blocking rate and the energy consumption of all the available flickering patterns, which are not available. We therefore are motivated to propose an alternative approach which only requires the information available.

The MM selection and update are carried out periodically. It is reasonable to consider a control delay until the selected MM is reflected in the system. In this paper, the selection cycle and the control delay are assumed to be the same, which is denoted by  $\Delta t$  (transmission cycles). We denote the selected MM for transmission cycle  $t$  by  $\mathbf{M}_{m(t)} \in \mathcal{M}$ , which will be used during transmission cycle  $[t + \Delta t, t + 2\Delta t)$ . We call  $m(t)$  the MM index (MMI)<sup>3</sup>. The current traffic load  $\rho_b(t, m(t), s)$  and blocking rate  $P_{\text{blk}}(t, m(t))$  are available for MM selection. In the following, without loss of generality, we consider the update during transmission cycle  $t$ .

<sup>3</sup>We have included  $t$  as a variable so as to indicate the different MM at each cycle.

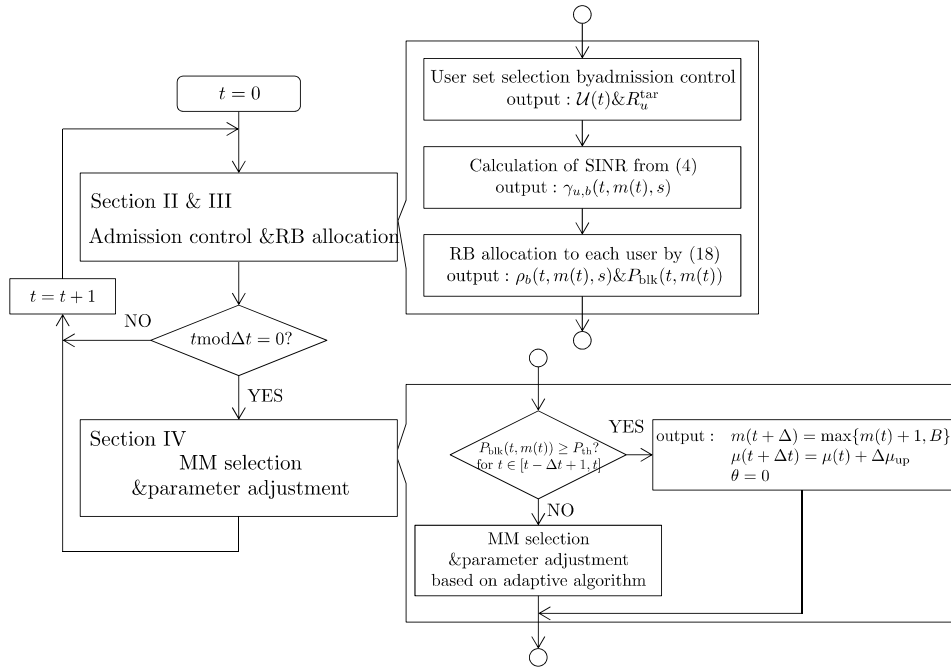


Fig. 4. Overall procedure of the proposed MM selection algorithm.

#### A. Approximation of Traffic Load for MM Selection

Let us define the average traffic load of the cluster with  $\mathbf{M}_{m(t)}$  as

$$\bar{\rho}(t, m(t)) \triangleq \frac{1}{Bm(t)} \sum_{b \in \mathcal{B}} \sum_{s=1}^S \rho_b(t, m(t), s), \quad (28)$$

where  $\rho_b(t, m(t), s)$  is obtained by solving (18). Although the traffic demand such as the number of users and QoS requirement exhibits a time-varying nature, *high correlation is expected within short time interval*, i.e.,  $\bar{\rho}_b(t + \Delta t, m(t)) \approx \bar{\rho}_b(t, m(t))$  if the same  $\mathbf{M}_{m(t)}$  is used.

To enable the MM selection, we need to predict the traffic load with different MM other than  $\mathbf{M}_{m(t)}$  as

$$\tilde{\rho}(t + \Delta t, m(t) - 1) \approx \tilde{\rho}(t, m(t) - 1) = f(\bar{\rho}(t, m(t))), \quad (29)$$

where  $f(\cdot)$  is a function. To find the optimum  $f(\cdot)$  is quite difficult if it is not impossible. Fig. 3 shows the probability density function (pdf) of  $\bar{\mu} \triangleq \frac{\bar{\rho}(t, m(t) - 1)}{\bar{\rho}(t, m(t))}$  with  $B = 4$  when target rate of each user is randomly selected from target rate set  $\mathcal{R}^{\text{tar}} = \{0.1, 0.2, \dots, 1.0\}$  Mbps. From the figure, it can be seen that  $\bar{\mu}$  concentrates at certain value for different  $m(t)$ . This observation motivates us to *estimate*  $\bar{\rho}(t, m(t) - 1)$  from actual traffic load  $\bar{\rho}(t, m(t))$  by the following simple yet effective function:

$$\tilde{\rho}(t + \Delta t, m(t) - 1) \approx \mu(t) \bar{\rho}(t, m(t)), \quad (30)$$

where  $\mu(t)$  is the *multiplication factor*. **C1** will be satisfied if

$$\tilde{\rho}(t + \Delta t, m(t) - 1) \leq 1. \quad (31)$$

Then the total energy consumption of the cluster with

$\mathbf{M}_{m(t)-1}$  is estimated as

$$\begin{aligned} \tilde{E}_c(t + \Delta t, m(t) - 1) \\ = t_{\text{ts}} B ((m(t) - 1) P(\tilde{\rho}(t + \Delta t, m(t) - 1) \\ + (S - (m(t) - 1))) P_{\text{NM}}). \end{aligned} \quad (32)$$

**C2** is considered to be satisfied if

$$\tilde{E}_c(t + \Delta t, m(t) - 1) < E_c(t, m(t)). \quad (33)$$

#### B. Adaptation of MM

Fig. 4 summarizes the overall procedure of the proposed RB allocation and MM selection algorithm. Every  $\Delta t$  transmission cycles,  $m(t)$  and  $\mu(t)$  are updated. The decision made at transmission cycle  $t$  is reflected on the operation for transmission cycle  $[t + \Delta t, t + 2\Delta t)$ . There are three transition cases of the MMI as shown in Fig. 5:

- $\text{Tr}_{m(t)}^0$ :  $m(t + \Delta t) = m(t)$ ,
- $\text{Tr}_{m(t)}^+$ :  $m(t + \Delta t) = \min\{m(t) + 1, B\}$ , and
- $\text{Tr}_{m(t)}^-$ :  $m(t + \Delta t) = \max\{m(t) - 1, 1\}$ ,

where  $\min\{a, b\}$  and  $\max\{a, b\}$  return the minimum and the maximum of  $a$  and  $b$ , respectively. If  $P_{\text{blk}}(t, m(t)) > P_{\text{th}}$ , we then need to change the MMI in order to support the current traffic demand. The violation of **C1** is mainly due to the reduced number of TM time slots. Hence, we update the MMI as  $m(t + \Delta t) = \min\{m(t) + 1, B\}$  by the transition  $\text{Tr}_{m(t)}^+$ . On the other hand, if  $P_{\text{blk}}(t, m(t)) \leq P_{\text{th}}$ , the current traffic demand can be supported by  $\mathbf{M}_{m(t)}$ .  $\tilde{\rho}(t + \Delta t, m(t) - 1)$  is used to check if we can achieve energy saving without the violation of **C1** and **C2** by the transition  $\text{Tr}_{m(t)}^-$ . However, as  $\tilde{\rho}(t + \Delta t, m(t) - 1)$  is the estimate of  $\bar{\rho}(t + \Delta t, m(t) - 1)$ , incorrect decisions may be made as follows:

- False alarm (FA): Transition  $\text{Tr}_{m(t)}^-$  is triggered, but either **C1** or **C2** is not satisfied.



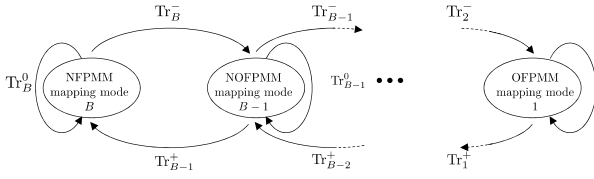


Fig. 5. State transition among available MM.

- Miss detection (MD): Transition  $\text{Tr}_{m(t)}^0$  is triggered, but **C1** and **C2** are satisfied.

FA happens if the estimated traffic load is lower than the actual value, while MD happens if the estimated traffic load is higher than the actual one.

Correspondingly, if  $\mu(t)$ , which is used to estimate the traffic load, is smaller than the actual value, the traffic load is *underestimated* and FA is incurred; if  $\mu(t)$  is greater than the actual value, the traffic load is *overestimated* and this results in MD. Therefore, a *fixed*  $\mu(t)$  is not appropriate. These observations motivate us to adaptively adjust  $\mu(t)$  based on the time-varying traffic demand as follows. To adjust  $\mu(t)$ , we introduce counter  $\theta$  and its threshold  $\Theta$ .  $\theta$  is incremented if **C1** is not violated since the last update and is set to zero if **C1** is violated.

By setting the threshold value  $\Theta$  higher, the algorithm becomes more *conservative*, i.e., it places more importance on the QoS degradation than energy saving. On the other hand, if the threshold value  $\Theta$  is smaller, the algorithm emphasizes energy saving, i.e., it becomes *aggressive*.

- If **C1** is not violated for a period of time, i.e.,  $\theta = \Theta$ ,  $\mu(t + \Delta t) = \mu(t) - \Delta\mu_{\text{down}}$ , where  $\Delta\mu_{\text{down}}$  is the predetermined step size. MD can be prevented.
- If **C1** is violated,  $\mu(t + \Delta t) = \mu(t) + \Delta\mu_{\text{up}}$ , where  $\Delta\mu_{\text{up}}$  is the predetermined step size. FA can be prevented.

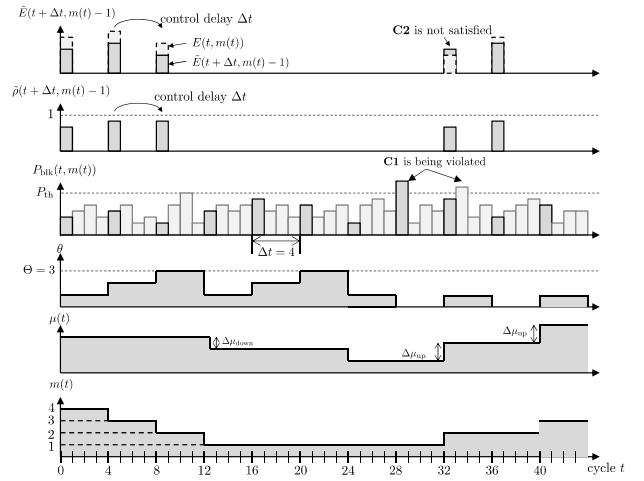
As it is summarized in Table II, the MMI update is done as

$$m(t + \Delta t) = \begin{cases} \min\{m(t) + 1, B\} & \text{if } P_{\text{blk}}(t, m(t)) > P_{\text{th}}, \\ \max\{m(t) - 1, 1\} & \text{else if (31) and (33) are satisfied,} \\ m(t) & \text{otherwise.} \end{cases} \quad (34)$$

During the initialization step ( $t = 0$ ), we set  $\mu(t) = \mu_{\text{init}}$ ,  $\theta = 0$ , and  $m(0) = B$ .

During an execution step ( $t \geq 0$ ), RBs are allocated to users by solving (18) with  $\mathbf{M}_{m(t)}$  and  $P_{\text{blk}}(t, m(t))$  is calculated and stored for next update. At every  $\Delta t$  transmission cycles, the following updates are carried out.

- If  $P_{\text{blk}}(t, m(t)) > P_b^{\text{th}}$  for  $t \in [t - \Delta t + 1, t]$ , **C1** is violated since the last update. Thus, the number of available time slots is increased by the transition  $\text{Tr}_{m(t)}^+$ . To prevent FA,  $\mu(t + \Delta t) = \mu(t) + \Delta\mu_{\text{up}}$  and  $\theta$  is reset to zero.
- If  $P_{\text{blk}}(t, m(t)) \leq P_b^{\text{th}}$  for  $t \in [t - \Delta t + 1, t]$ ,  $\tilde{p}(t + \Delta t, m(t) - 1)$  and  $\tilde{E}_c(t + \Delta t, m(t) - 1)$  are estimated by (30) and (32) based on  $\tilde{p}(t, m(t))$ , respectively, and  $\theta = \theta + 1$ . If and only if  $\tilde{p}(t + \Delta t, \max\{m(t) - 1, 1\}) \leq 1$  and  $\tilde{E}_c(t + \Delta t, m(t) - 1) < E_c(t, m(t))$ , we set  $m(t + \Delta t) = m(t) - 1$ . Otherwise  $m(t + \Delta t) = m(t)$ . If  $\theta = \Theta$ ,

Fig. 6. An example of adaptive MM selection with  $\Delta t = 4$  and  $\Theta = 3$ .

$\mu(t + \Delta t) = \mu(t) - \Delta\mu_{\text{down}}$  and reset the counter  $\theta = 0$ . Otherwise,  $\mu(t + \Delta t) = \mu(t)$ . Note that  $\theta$  only affects the decrease of  $\mu(t)$ .

Fig. 6 illustrates an example of operation of the proposed MM selection algorithm with  $\Delta t = 4$  and  $\Theta = 3$ . In this example,  $m(t)$  is decreased every  $\Delta t = 4$  cycles by the transition  $\text{Tr}_{m(t)}^-$  because **C1** is satisfied and **C1** and **C2** would be satisfied based on the estimates.  $\mu(t)$  is decreased by  $\Delta\mu_{\text{down}}$  at cycles 12 and 24 as  $\theta = \Theta$ . The transition  $\text{Tr}_1^+$  is triggered during cycle 28 due to the violation of **C1**. To avoid FA,  $\mu(t)$  is increased by  $\Delta\mu_{\text{up}}$  and  $\theta = 0$ . In the update during cycle 32, only **C1** is satisfied. Thus, the transition  $\text{Tr}_2^0$  is triggered. During the update in cycle 36, the transition  $\text{Tr}_2^+$  is triggered due to the violation of **C1** since last update, i.e., at cycle 33.

### C. Implementation

To implement the adaptive CoNap in a practical system, the system requirements such as a control delay should be taken into account. Since the flickering pattern of CoNap is flexible, an arbitrary flickering pattern other than G-FPM is possible to satisfy the system requirement. Furthermore, there is *no need to explicitly* inform the users the flickering pattern of each BS. The current status of the flickering pattern of each BS can be implicitly informed to the users. Since the mobile station (MS) needs to monitor the control channel in the current cellular system, the MS can detect when each BS is in TM or NM by detecting the presence of the control signal. Thus, no additional signaling is necessary.

The MM selection is performed every  $\Delta t$  cycles (selection cycle). The MM selection cycle is system dependent. If the time-varying traffic demand needs to be tracked closely, the selection cycle may be reduced to, e.g., less than several seconds. The centralized approach is most promising for MM selection. In practical systems, each BS is connected to an entity which controls the network. This entity may take the role of a central unit (CU) to perform MM selection based on the information provided by the BSs within a cluster. In such a case, the *current blocking rate* and the *current traffic load* information need to be sent from each BS to CU for MM selection. MMI is then fed back to each BS. This



TABLE II  
TWO CONDITIONS AND MMI TRANSITIONS

| MMI<br>during cycle $t$ | cycle $t$ (actual)<br>C1 | cycle $t + \Delta t$ (estimated)<br>C1 C2 |   | Transition           | MMI<br>during cycle $[t + \Delta t, t + 2\Delta t)$ |
|-------------------------|--------------------------|-------------------------------------------|---|----------------------|-----------------------------------------------------|
| $m(t)$                  | ×                        | —                                         | — | $\text{Tr}_{m(t)}^+$ | $\min\{m(t) + 1, B\}$                               |
|                         | ○                        | ○                                         | ○ | $\text{Tr}_{m(t)}^-$ | $\max\{m(t) - 1, 1\}$                               |
|                         | ○                        | ○                                         | × | $\text{Tr}_{m(t)}^0$ | $m(t)$                                              |
|                         | ○                        | ×                                         | ○ |                      |                                                     |
|                         | ○                        | ×                                         | × |                      |                                                     |

information is significantly small compared to the user data or CSI which is required for other coordination approaches. As for the complexity, any scheduling algorithm can be adopted in the proposed adaptive CoNap. This means that the additional complexity incurred by the application of the proposed scheme is solely due to the MM selection. However, the selection algorithm itself can be implemented with low complexity.

## V. SIMULATION RESULTS

In this section, the energy consumption and the blocking rate of the adaptive proposed CoNap are evaluated. The simulation parameters are summarized in Table III. A simple admission control is considered in the evaluation. At each transmission period  $t$ , a user set is generated according to the predetermined number of users within a cluster. The users are uniformly and randomly distributed within a cluster. Association of each user with BS and RB allocation are performed by the algorithm developed in Section III. A with N-FPMM, i.e.,  $m = B$ . If  $P_{\text{blk}}(t, B) > 0.1$ , a new user set is generated within the cluster. The procedure is repeated until  $P_{\text{blk}}(t, B) \leq 0.1$ . Each user is equipped with two receive antennas. The simulated network consists of 28 cells. The cluster size is set to  $B = 4$  as it compromises the energy saving gain and the increase of blocking rate [26]. Thus, there are seven clusters. The performance is evaluated in the central cluster and the interferences from surrounding first tier clusters are taken into account. The distance between neighboring BSs is set to 500m. We consider a distance dependent path loss at the carrier frequency of 2GHz and it is modeled as  $L_{u,b} = 128.1 + 37.6 \log_{10}(d_{u,b})$  where  $d_{u,b}$  is the distance between user  $u$  and BS  $b$  in km and a log-normally distributed shadowing with the standard deviation of  $\sigma = 8\text{dB}$  [28]. Antenna gain  $A(\psi_{u,b})$  is calculated as  $A(\psi_{u,b}) = -\min[12(\frac{\psi_{u,b} - \psi_{\text{etilt}}}{\psi_{3\text{dB}}}), A_v]$ , where  $\psi_{u,b}$  is the relative elevation angle of the user  $u$  from the main-beam direction of the antenna of BS  $b$ ,  $\psi_{3\text{dB}}$  is the half power beamwidth (HPBW),  $\psi_{\text{etilt}}$  is the electrical antenna downtilt, and  $A_v$  is the sidelobe floor [28]. The height of BS antenna and user is set to 32m and 1.5m, respectively [28]. The target rate of users  $R_u^{\text{tar}}$  is randomly selected from the target rate set  $\mathcal{R}^{\text{tar}} = \{0.1, 0.2, \dots, 1.0\}\text{Mbps}$ . One transmission cycle  $t_{\text{cycle}}$  is set to 1sec. Each value of the power consumption in (15) is given as  $P_{\text{fix}} = 10.8\text{Watt}$ ,  $P_{\text{dyn}} = 14.8\text{Watt}$ , and  $p_{\text{loss}} = 1.257$  [2]. For adaptive selection algorithm, the initial value and the step sizes of  $\mu(t)$  are set to  $\mu_{\text{init}} = 0.5$ ,  $\Delta\mu_{\text{up}} = 0.5$ , and  $\Delta\mu_{\text{down}} = 0.1$ .

TABLE III  
SIMULATION PARAMETERS

| Cell layout                    |                       |               |
|--------------------------------|-----------------------|---------------|
| Cluster size                   | $B$                   | 4             |
| Inter-BS distance              | $r_{\text{cell}}$     | 500 (m)       |
| System Parameters              |                       |               |
| Total bandwidth                | $W_{\text{total}}$    | 10 (MHz)      |
| Number of frequency domain RBs | $N$                   | 50            |
| Transmission cycle             | $t_{\text{cycle}}$    | 1 (sec)       |
| Maximum transmit power         | $P_{\text{max}}$      | 46 (dBm)      |
| Overhead of radiated power     | $p_{\text{OH}}$       | 0.1           |
| OBO                            | $\xi$                 | 8 (dB)        |
| Channel & Antenna Parameters   |                       |               |
| Shadowing standard deviation   | $\sigma$              | 8 (dB)        |
| AWGN power density             | $\sigma_n^2$          | -174 (dBm/Hz) |
| HPBW                           | $\psi_{3\text{dB}}$   | 10 (degrees)  |
| Sidelobe floor                 | $A_v$                 | 20 (dB)       |
| Electrical antenna downtilt    | $\psi_{\text{etilt}}$ | 15 (degrees)  |

### A. Impact of FPMMs Under Fixed Number of Users

We evaluate the impact of FPMM on the BS energy consumption. For comparison, the system with N-FPMM but without solving (18), which is denoted as “N-FPMM w/ random scheduling”, is considered. In this scheme, BS allocates the RBs to its corresponding users at randomly selected time slot. Note that solving (18) for N-FPMM is equivalent to “on-off scheduler” [2]. The flickering pattern of O-FPMM is equivalent to CS-O [15]. For fair performance comparison, we use the same RB allocation algorithm in Sect. III.A for CS-O. Fig. 7 shows that the CS-O provides the lowest energy consumption, followed by the G-FPMMs ( $m = 2$  and 3), due to the ICI reduction. By setting  $m = 2$  for G-FPMM, further energy saving is achieved compared to G-FPMM with  $m = 3$ . This is because the BSs are in NM for larger number of time slots. However, as shown in Table IV, energy saving is obtained at the cost of the blocking rate increase, i.e., network QoS violation. The CS-O provides the significant energy saving without compromising network QoS when the number of users is large (corresponds to traffic demand). Similarly, the G-FPMMs can achieve the energy saving even when the number of users is large, i.e.,  $|U/B| = 30$ .

### B. Time-Varying Traffic Demand

The performance of the proposed adaptive CoNap under the time-varying traffic demand is evaluated. The time-varying traffic demand is characterized by the total number of users in the cluster, where the number of users is assumed to be same for  $T$  consecutive transmission cycles. In this paper, we set  $T = 60$ . Thus, the number of users within a cluster is fixed during 1min but their QoS requirement, location, and channel

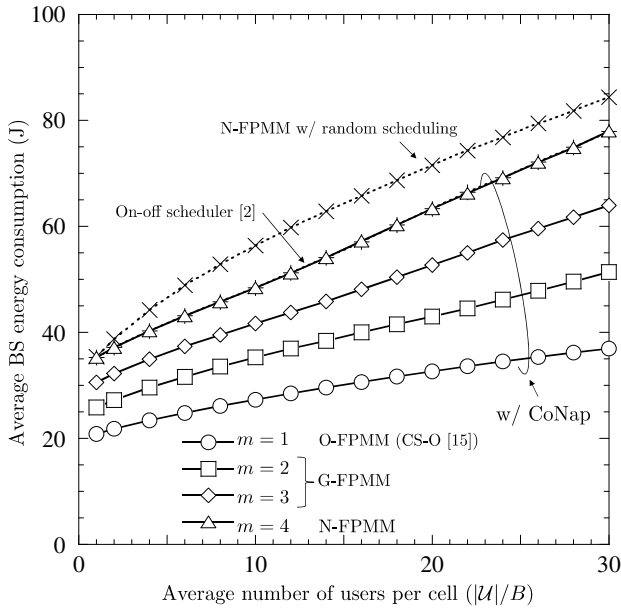


Fig. 7. The impact of various FPMMs on the energy consumption.

condition are independently generated during each simulation run of 1sec. The number of users per cluster during cycle  $t \in [t'T, (t' + 1)T)$  is calculated as

$$|U(t')| = B \left[ U_{\text{mean}} + U_{\text{var}} \cos \left( 2\pi \frac{t' + t_{\text{offset}}}{1440} \right) + 2v \right], \quad \text{for } 1 \leq t' \leq 1440, \quad (35)$$

where  $U_{\text{mean}}$  and  $U_{\text{var}}$  are the average number and the variance of users per cluster, respectively, and  $v \sim \mathcal{N}(0, 1)$  is a Gaussian random variable. In Fig. 8, we set  $U_{\text{mean}} = 20$ ,  $U_{\text{var}} = 15$ , and  $t_{\text{offset}} = 600$ .

For MM selection, two MM sets are considered:  $\mathcal{M}_1 = \{\mathbf{M}_1, \mathbf{M}_4\}$  and  $\mathcal{M}_2 = \{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$ . We set  $\Delta t = 10$  and  $\Theta = 60$ . The BS energy consumption is shown in Fig. 9. For comparison, the BS energy consumptions of on-off scheduler [2] and CS-O [15] are also included in the figure. The on-off scheduler provides slight energy saving gain while CS-O provides significant energy saving gain due to the small power consumption during NM and the ICI reduction. However, the energy saving of CS-O is at the sacrifice of QoS as it will be shown later in Fig. 11. During the low traffic demand condition, the adaptive CoNap with different MM sets provides the same energy consumption as CS-O. This is because the O-FPMM is always selected, i.e.,  $m(t) = 1$ . During relatively high traffic demand condition, however, the choice of MM set results in different energy consumptions. This can be explained as follows. As the traffic demand increases, the network QoS requirement cannot be supported by O-FPMM as shown in Table IV. Thus the algorithm tries to use the other MMs. However, the remaining MM in the set  $\mathcal{M}_1$  is the N-FPMM, whereas there are other G-FPMMs in the set  $\mathcal{M}_2$ . Therefore the set  $\mathcal{M}_2$  can achieve further energy saving gain during high traffic demand condition.

Fig. 10 shows the accumulated BS energy consumption for one day. We set  $\Delta = 10$  and  $\Theta \in \{10, 60\}$ . For comparison,

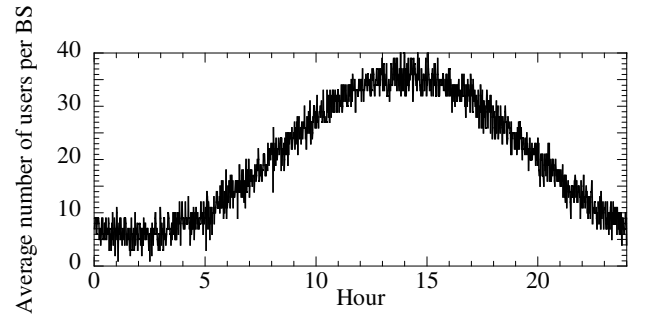


Fig. 8. Time varying traffic demand.

TABLE IV  
AVERAGE BLOCKING RATE OF ON-OFF SCHEDULER [2], CS-O [15], AND CoNap WITH G-FPMM WHEN TARGET RATE OF USERS ARE RANDOMLY SELECTED FROM  $\mathcal{R}^{\text{tar}} = \{0.1, 0.2, \dots, 1.0\}$  MBPS.

| $ U /B$ | on-off scheduler     | CS-O                 | w/ CoNap             |                      |
|---------|----------------------|----------------------|----------------------|----------------------|
|         |                      |                      | $m = 2$              | $m = 3$              |
| 10      | 0                    | 0                    | 0                    | 0                    |
| 20      | 0                    | $5 \cdot 10^{-6}$    | $1 \cdot 10^{-5}$    | $2.5 \cdot 10^{-6}$  |
| 30      | $1.25 \cdot 10^{-4}$ | $9.24 \cdot 10^{-3}$ | $5.85 \cdot 10^{-4}$ | $7.67 \cdot 10^{-5}$ |

the adaptive CoNap with ideal MM selection (denoted by “*Ideal CoNap*”) is considered. We consider two criteria for ideal CoNap; *conservative* and *aggressive*. For the conservative criterion, the MM which provides the lowest blocking rate is selected. Whereas, for the aggressive criterion, the MM which provides the minimum energy consumption without violation of **C1** is selected. The figure shows that the proposed adaptive algorithm with  $\Theta = 60$  becomes more conservative than that with  $\Theta = 10$ . By setting  $\Theta = 10$ , it becomes more aggressive and provides the energy consumption close to ideal CoNap with aggressive criterion. The adaptive CoNap can save up to 40% energy compared to the conventional system (on-off scheduler) after 24 hours.

The proposed adaptive CoNap reduces the number of available time slot for each BS. Therefore user QoS, i.e., target rate, may not be satisfied, however, several RBs may be allocated to users through RB allocation. Let us define the variable  $\epsilon_u \triangleq \frac{R_u}{R_u^{\text{tar}}}$ , which is the total rate the user obtained divided by its target rate. If  $\epsilon_u < 1$ , the user is considered to be blocked. Otherwise, the target rate of user  $u$  is satisfied. Fig. 11 shows the cumulative distribution function (CDF) of  $\epsilon_u$ . Since  $\epsilon_u$  is the same for “N-FPMM w/ random scheduling” and “on-off scheduler” [2], only one line is shown in the figure. For the proposed adaptive CoNap, by increasing  $\theta$ , the algorithm becomes *conservative* to emphasize more QoS guarantee than energy saving, hence the probability of user QoS being satisfied increases by increasing  $\Theta$  as the algorithm becomes *conservative*, i.e., it gives higher importance to QoS guarantee than energy saving. More than 99.5 (99.8)% of the users can satisfy their target rate by setting  $\Theta = 10$  (60). Although significant energy saving is obtained by CS-O [15], only about 96.9% of users satisfy their QoS. On the other hand, the on-off scheduler [2] guarantees about 99.9% QoS requirement. Thus, we can conclude that the proposed adaptive CoNap with MM selection method well balances the energy saving with the QoS guarantee.

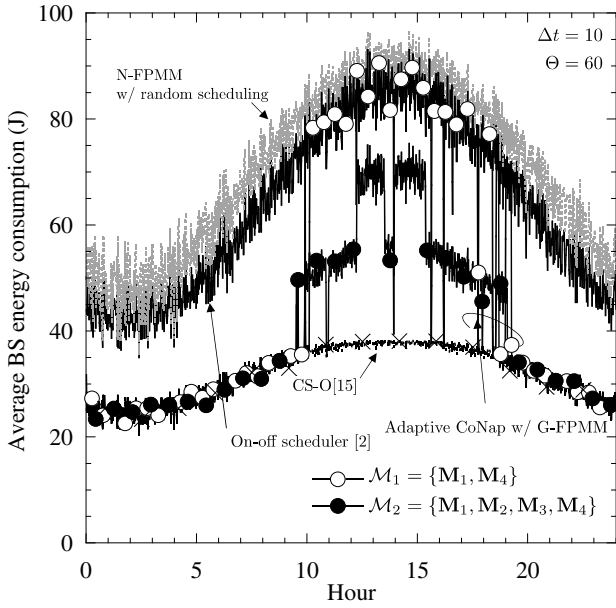


Fig. 9. The total energy consumption of the adaptive CoNap with  $\Delta t = 10$  and  $\Theta = 60$  under time-varying traffic demand.

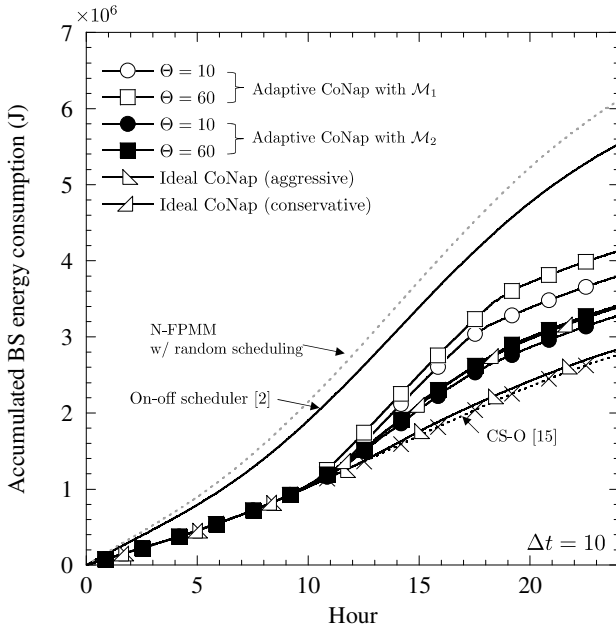


Fig. 10. Accumulated BS energy consumption for one day with  $\Delta t = 10$ .

## VI. CONCLUSION

We have proposed an adaptive coordinated napping (CoNap) to save the energy consumption of wireless communication system. By forming a cluster consisting of multiple BSs, the implicit coordination among BSs is realized by a binary general flickering pattern matrix (FPM) and a mapping matrix (MM). To track the time-varying traffic demand of the system and enable implicit multi-cell (BS) coordination, an adaptive MM selection algorithm is developed. The MM which achieves the energy gain while satisfying the QoS requirement is selected from the set of available MMs based on the traffic load. Through computer simulation, it is shown that the adaptive CoNap can significantly reduce the energy

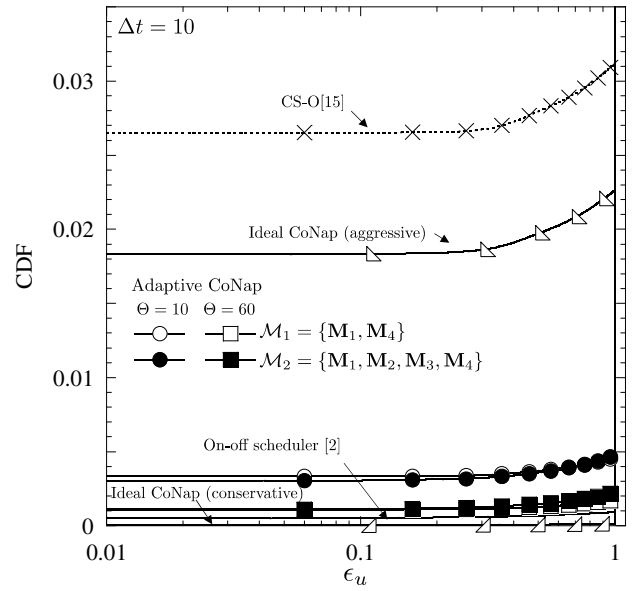


Fig. 11. CDF of  $\epsilon_u$ . If  $\epsilon_u = 1$ , the user QoS is satisfied, and otherwise, outage happens.

consumption, i.e., around 40% saving compared to the conventional system without napping under the time-varying traffic demand. The proposed adaptive CoNap does not require any explicit coordination among the BSs, any data sharing, and exchange of CSI; thus, the system complexity and backhaul network overhead can be also reduced.

## APPENDIX

### DERIVATION OF DOHERTY PAE (13a)

From [19, (10.18)], we have

$$\eta = \frac{\pi}{2} \left( \frac{v_{in}}{V_{max}} \right), \quad \text{for } 0 < \frac{v_{in}}{V_{max}} < \frac{1}{2}, \quad (36)$$

where  $v_{in}$  is the input voltage and  $V_{max}$  is the maximum (peak) input voltage to PA. Since  $V = \sqrt{PR_{opt}}$  with  $R_{opt}$  being the resistor of the PA, we have

$$\begin{cases} v_{in} &= \sqrt{P_{RF}(\rho_b(m, s))R_{opt}}, \\ V_{max} &= \sqrt{\xi P_{max} R_{opt}}. \end{cases} \quad (37)$$

By substituting (37) into (36), we have

$$\eta = \frac{\pi}{2} \sqrt{c P_{RF}(\rho_b(m, s))}, \quad \text{for } 0 < c P_{RF}(\rho_b(m, s)) < \frac{1}{4}, \quad (38)$$

(13b) can be derived in the similar way.

## REFERENCES

- [1] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. G3dor, G. Auer, and L. V. der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, pp. 66–72, Nov. 2010.
- [2] INFISO-ICT-247733 EARTH Deliverable D2.3, Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. Available: <https://www.ict-earth.eu>.
- [3] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.



- [4] J. Joung, C. K. Ho, and S. Sun, "Tradeoff of spectral and energy efficiencies: impact of power amplifier on OFDM systems," in *Proc. 2012 IEEE Global Commun. Conf.*, pp. 3274–3279.
- [5] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Commun. Mag.*, vol. 47, no. 3, pp. 88–95, Mar. 2009.
- [6] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: a survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, pp. 524–540, Forth quarter 2011.
- [7] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. 2009 GreenComm — Int. Workshop Green Commun.*, pp. 1–5.
- [8] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, pp. 74–79, Nov. 2010.
- [9] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proc. 2010 IEEE Global Commun. Conf.*, pp. 1–5.
- [10] S. Han, C. Yang, G. Wang, and M. Lei, "On the energy efficiency of base station sleeping with multicell cooperative transmission," in *Proc. 2011 IEEE Int. Symp. Personal, Indoor Mobile Radio Commun.*, pp. 1536–1540.
- [11] Z. Niu, S. Zhou, Y. Hua, Q. Zhang, and D. Cao, "Energy-aware network planning for wireless cellular system with inter-cell cooperation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1412–1423, Apr. 2012.
- [12] P. Yu, Q. Yang, F. Fu, and K. S. Kwak, "Inter-cell cooperation aided dynamic base station switching for energy efficient cellular networks," in *Proc. 2012 Asia-Pacific Conf. Commun.*, pp. 159–163.
- [13] A. Bousia, E. Kartsakli, L. Alonso, and C. Verikoukis, "Dynamic energy efficient distance-aware base station switch on/off scheme for LTE-Advanced," in *Proc. 2012 IEEE Global Commun. Conf.*, pp. 1532–1537.
- [14] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Gódor, "Reducing energy consumption in LTE with cell DTX," in *Proc. 2011 IEEE Veh. Technol. Conf. — Spring*, pp. 1–5.
- [15] K. Abdallah, I. Cerutti, and P. Castoldi, "Energy-efficient coordinated sleep of LTE cells," in *Proc. 2012 IEEE Int. Conf. Commun.*, pp. 6760–6764.
- [16] J. Wu, Y. Wu, S. Zhou, and Z. Niu, "Traffic-aware power adaptation and base station sleep control for energy-delay tradeoffs in green cellular networks," in *Proc. 2012 IEEE Global Commun. Conf.*, pp. 3171–3176.
- [17] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedge, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 10–22, June 2010.
- [18] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, pp. 102–111, Feb. 2011.
- [19] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*, 2nd ed. Artech House, 2006.
- [20] J. Joung, C. K. Ho, and S. Sun, "Spectral efficiency and energy efficiency of OFDM systems: Impact of power amplifiers and countermeasures," *IEEE J. Sel. Areas Commun.*, vol. 32, pp. 1–13, Dec. 2014 (IEEE early access articles).
- [21] R. Zakhour and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Trans. Signal Process.*, vol. 12, no. 59, pp. 6102–6111, Dec. 2011.
- [22] B. Rengarajan, G. Rizzo, and M. A. Marsan, "Bounds on QoS-constrained energy savings in cellular access networks with sleep modes," in *Proc. 2011 Int. Teletraffic Congress*, pp. 47–54.
- [23] H. Bogucka and A. Conti, "Degrees of freedom for energy savings in practical adaptive wireless systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 38–45, June 2011.
- [24] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, June 2011.
- [25] G. Auer, V. Giannini, M. Olsson, M. J. Gonzalez, and C. Desset, "Framework for energy efficiency analysis of wireless networks," in *Proc. 2011 Int. Conf. Wireless Commun., Veh. Technol., Inf. Theory Aersp. Electron. Syst. Technol.*, pp. 1–5.
- [26] K. Adachi, J. Joung, S. Sun, and P. H. Tan, "Energy-saving coordinated napping (CoNap) for wireless networks," in *Proc. 2012 IEEE Global Commun. Conf., Workshop Multi-cell Cooperation*, pp. 1150–1154.
- [27] F.-S. Chu, K.-C. Chen, and G. Fettweis, "Green resource allocation

to minimize receiving energy in OFDMA cellular systems," *IEEE Commun. Lett.*, vol. 16, no. 3, pp. 372–374, Mar. 2012.

- [28] 3GPP TR 36.814 V9.0.0, "Further advancements for E-UTRA physical layer aspects," 2010.



**Koichi Adachi (M'10)** received the B.E., M.E., and Ph.D degrees in engineering from Keio University, Japan, in 2005, 2007, and 2009 respectively. From 2007 to 2010, he was a Japan Society for the Promotion of Science (JSPS) research fellow. Currently he is with the Institute for Infocomm Research, A\*STAR, in Singapore. His research interests include cooperative communications. He was the visiting researcher at City University of Hong Kong in April 2009 and the visiting research fellow at University of Kent from June to Aug 2009.

Dr. Adachi served as General Co-chair of the Tenth IEEE Vehicular Technology Society Asia Pacific Wireless Communications Symposium (AP-WCS) and Track Co-chair of Transmission Technologies and Communication Theory of the 78th IEEE Vehicular Technology Conference in 2013. He was recognized as the Exemplary Reviewer from IEEE COMMUNICATIONS LETTERS and IEEE WIRELESS COMMUNICATIONS LETTERS in 2012.



**Jingon Joung (M'07)** received B.S. degree in EE from Yonsei University, Korea, in 2001, and the M.S. and Ph.D. degrees in EECS from the Korea Advanced Science and Technology (KAIST), Korea, in 2003 and 2007, respectively. He was a Postdoctoral Research Scientist at EE Department, KAIST, from March 2007 to August 2008, and was a Postdoctoral Fellow at EE Department, University of California, Los Angeles, USA, from October 2008 to September 2009. Since October 2009, he has been working as a Scientist at Institute for Infocomm Research,

Singapore. His current research has focused on the study of energy efficient wireless communications including multiuser-, large-, and cooperative-MIMO techniques.

Dr. Joung was recognized as the Exemplary Reviewers in 2012 from IEEE COMMUNICATIONS LETTERS and IEEE WIRELESS COMMUNICATIONS LETTERS, and he was the recipient of the first prize at the Intel-ITRC Student Paper contest in 2006.



**Sumei Sun (SM'12)** received the B.Sc. (with honors) degree from Peking University, Beijing, China; the M.Eng. degree from Nanyang Technological University, Singapore; and the Ph.D. degree from National University of Singapore, Singapore.

She has been with Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A\*STAR), Singapore, since 1995, where she is currently Head of the Modulation and Coding Department, developing energy- and spectrum-efficient technologies for the next-generation communication systems. Her recent research interests include 5G transmission technologies, renewable energy management and cooperation in wireless systems and networks, and wireless transceiver design.

Dr. Sun served as the Technical Program Committee Chair of the 12th IEEE International Conference on Communications Systems (ICCS) in 2010, General Co-chair of the Seventh and Eighth IEEE Vehicular Technology Society Asia Pacific Wireless Communications Symposium, and Track Co-chair of Transmission Technologies of the 75th IEEE Vehicular Technology Conference in 2012. She is also an Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Editor for IEEE WIRELESS COMMUNICATIONS LETTERS. She was a co-recipient of the 16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications Best Paper Award.



**Peng Hui Tan (M'05)** received the M. Eng in the electrical and electronic engineering from National University of Singapore, Singapore and the Ph.D. degree from Chalmers University of Technology in 1999 and 2005, respectively. Now he is a Research Fellow at Institute for Infocomm Research. His areas of research are coding in relay network, scheduling, multiuser detection and iterative decoding.