Energy-efficient Design of Heterogeneous Cellular Networks from Deployment to Operation

Kyuho Son[a], Eunsung Oh[b,*], Bhaskar Krishnamachari[c]

[a]*T-Mobile US, Bellevue, WA 98006*
[b]*Department of Electronics Engineering, Hanseo University, South Korea 356-706*
[c]*Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089*

Abstract

The ever-increasing traffic demand has motivated mobile operators to explore how they can boost their network capacity with a minimal increase in their capital and operating expenditures. In order to tackle this problem, we investigate the energy-efficient design of heterogeneous cellular network (or simply HetNet), especially with a focus on deployment and operation strategies. We first formulate a general problem pertaining to minimizing the total energy consumption cost while satisfying the requirement of area spectral efficiency (ASE). We decompose this problem into a deployment problem at peak time and an operation problem at off-peak time. Under practical assumptions made from an observation on various topologies including an acquired real base-station deployment dataset, we demonstrate the submodularity of ASE function with respect to micro base-station deployment. Subsequently, we propose a greedy algorithm that is shown to be a constant-factor approximation to the optimal deployment. Although the greedy algorithm can be applied as an offline centralized solution for the operation problem, we further propose two online distributed algorithms with low complexity and signaling overhead using Lagrangian relaxation technique. Extensive simulations show that the proposed algorithms can significantly reduce the energy consumption with minimal deployment of micro base-stations.

Keywords: Energy-efficient, heterogeneous cellular network (HetNet), deployment, operation, submodularity, greedy algorithm

1. Introduction

Motivated by the explosive traffic demand from bandwidth-hungry multimedia and Internet-related services in broadband cellular networks, communication network engineers seek to maximally exploit the spectral resources in all available dimensions. Heterogeneous cellular network (HetNet) [1–3] where small cells such as micro, pico

*Corresponding author. Tel.: +82 416601413
Email addresses: kyuho.son@t-mobile.com (Kyuho Son), esoh@hanseo.ac.kr (Eunsung Oh), bkrishna@usc.edu (Bhaskar Krishnamachari)

and femto are used as a way of additionally increasing capacity and coverage beyond the initial deployment of macro cells, has emerged as a promising solution. Many wireless standards such as 802.16m/WiMax2, 3GPP-LTE and LTE Advanced are also designed to support the heterogeneous cellular system. Incrementally deploying micro base stations (BSs) is much simpler than building out complex cell towers and macro BSs. Besides, it can also reduce both capital (e.g., hardware) and operating (e.g., electricity, backhaul and site lease) expenditures, which is especially attractive to mobile operators [4].

Meanwhile, *green networks* have recently received significant attention due to the depletion of non-renewable energy resources and a limit on CO_2 emissions. From the perspective of mobile operators, developing more energy-efficient networks is not only a matter of being green and responsible, but also an economically important issue. In particular, it is well known that BSs are one of the most energy-hungry segments in the cellular networks, which contributes to about 60-80% of the total energy consumption [5, 6]. However, they are often under-utilized such as at nighttime because being deployed by the operator targeting peak traffic usage. Even when a BS is experiencing little or no activity, it consumes the majority of its peak energy. For instance, a typical UMTS BS consumes between 800-1500W and has a transmission power of 20-40W for RF output. Therefore, beyond turning off only radio transceivers, dynamic approaches [6, 7] that allow the system to entirely switch off some under-utilized BSs and transfer the corresponding load to neighboring cells during low traffic period can substantially reduce the amount of wasted energy in the network.

1.1. Our Objective and Contributions

Although recent papers have made several steps towards green cellular networks[5–12], most studies have focused only on the network operation aspect. However, this paper considers both deployment and operation aspects in order to unburden the mobile operators from huge capital and operating expenditures (CAPEX & OPEX). To this end, our objective is to provide both theoretical framework and practical solutions for the following two key questions:

- *Deployment:* where and how many micro BSs need to be additionally deployed considering the traffic at peak time?
- *Operation:* how to operate (i.e., load-aware dynamic switching on-off) macro and micro BSs for energy conservation during off-peak times?

First, in the deployment problem, we try to find a minimal deployment of micro BSs while satisfying the requirement of area spectral efficiency (ASE), i.e., minimizing CAPEX. We make an observation from various topologies that there is a monotone relationship between coverage and ASE increment. Under such an assumption, we are able to prove the submodularity of the ASE function with respect to micro BS deployment and allows us to propose a greedy algorithm that can be shown to be a constant-factor approximation of optimal deployment. We also show through simulations that deploying micro BSs is much energy is inherently more energy-efficient than the conventional macro BSs.

Second, in the operation problem, our objective is to minimize energy consumption through dynamic BS operation, i.e., minimizing unnecessary OPEX. Even though the above greedy deployment algorithm can be also applied as a centralized offline solution for this problem, we further propose two distributed online algorithms using Lagrangian relaxation to have more practical solutions. Extensive simulations based on real cellular traffic traces and information regarding BS location that the proposed distributed algorithms not only can achieve the near performance of the centralized algorithm but also can significantly reduce the energy consumption by about 60-80% compared to the conventional static operation (i.e., always-ON strategy).

We would like to mention that this paper is in fact an extended version of our own prior work [1] that focused on the development of algorithms. In this paper, we strengthen our contributions by (i) further presenting technical analysis that make the proposed algorithms applicable to more general cases, (ii) justifying the effectiveness of two-step approach to tackle the original problem, and (iii) providing new theoretical results and proofs.

1.2. Related Work

A large body of research in HetNet has focused on resource allocation, e.g., spectrum allocation [13, 14], power control [15–17]; however, there has been relatively little work dealing with BS deployment. The studies in [18, 19] showed the energy consumption benefit of heterogeneous deployment only by simulations. In homogeneous setting (i.e., only one type of BSs), several BS deployment problems [20, 21] have been theoretically investigated. Stamatelos *et al.* [20] showed that an algorithm minimizing the overlapped coverage (i.e., the co-channel interference area) can maximize spectral efficiency in omni-antenna case. Srinivas *et al.* [21] proposed an algorithm which jointly considers both BS deployment and user assignment in backbone base mobile ad-hoc networks for throughput optimization. **In [22–24], the authors studied the energy efficiency and its relationship with the transmit powers and densities of macro/micro BSs in two-tier HetNet. In particular, [22] and [23] used a stochastic geometric based model to derive energy efficiency and area power consumption, respectively.** Shin *et al.* [25] proposed an iterative BS planning algorithm under simplified network models.

Our work differs from the previous works in that: (i) we present an analytical framework for optimal BS deployment in HetNet with the different types of BSs, and (ii) run extensive simulations based on BS topologies and traffic profiles acquired from real cellular networks.

Despite the fact that communication network engineers have been concerned with energy issues for decades, their main focus was to prolong battery life-time of mobile terminals or sensor nodes. Recently, there has been a shift of emphasis to the network-side as well but an amount of literature on green cellular networks is relatively scant compared to the mobile-side. Dynamic BS operation (i.e., switching-on/off BSs depending on the traffic profile) for energy saving has been investigated in [1, 5–9]. In [26], the authors considered to dynamically adapt the speed of computing system inside BSs for energy conservation. In addition, the concept of BS sharing, where different operators pool their BSs together to further conserve energy, was introduced in [5, 10].

However, most of the previous works [5, 6, 10] attempted to see how much energy saving can be possibly achieved under the deterministic traffic variation over time rather than developing algorithms that can be implemented in practice.

In several preliminary BS switching algorithms [7, 8], the authors do not capture the effect of the signal strength degradation when traffic loads are transferred from the switched-off BS to neighboring BSs. To reflect this effect, in this paper, (i) we consider a more sophisticated channel model based on signal to interference plus noise ratio (SINR), and (ii) propose practical and distributed algorithms for the dynamic BS operation.

The reminder of this paper is organized as follows. In Section 2, we formally describe our system model and general problem. In Sections 3 and 4, we propose deployment and operation algorithms, respectively. In Section 5, we demonstrate the performance of the proposed algorithms under various topologies and scenarios. Finally, we conclude the paper in Section 6.

2. System Description and Problem Definition

2.1. System Description

2.1.1. Network Model

Consider a heterogeneous cellular network where the sets of macro and micro BSs, denoted by $\mathcal{B}_M$ and $\mathcal{B}_m$, respectively. Those BSs lie in the two-dimensional area $\mathcal{A} \subset \mathbb{R}^2$. Let us denote by $b \in \mathcal{B} = \mathcal{B}_M \cup \mathcal{B}_m$ the index of BSs. Throughout the paper, subscript M is used for macro BSs, and m is for micro BSs. Our focus is on downlink communication as that is a primary usage mode for the mobile Internet, i.e., from BSs to mobile terminals (MTs). Although we focus on downlink communication, some aspects of our work can be applied to the uplink as well.

2.1.2. Link Model

The received signal strength from BS b to MT at location x can be expressed as $E_b(x) = p_b \cdot g_b(x)$, where p_b denotes the transmission power of BS b, $g_b(x)$ denotes the channel gain from BS b to location x, including path loss attenuation, shadowing and other factors if any. For analytical tractability, $g_b(x)$ is assumed not to change over time, but it can be considered as an time-averaged channel gain instead. This assumption is reasonable in the sense that the time scale of our problem for BS deployment and operation is much larger than the time scale of fast-fading.

Accordingly, SINR at location x can be written as[1]:

$$\Gamma(x, \mathcal{B}) = \frac{E_{b(x,\mathcal{B})}(x)}{\sum\limits_{b \in \mathcal{B} \setminus \{b(x,\mathcal{B})\}} E_b(x) + \sigma^2}, \tag{1}$$

[1] In (1), macro and micro BSs are basically considered to use the same frequency band (i.e., spectrum sharing). We can also incorporate the case where macro and micro BSs use different bands (i.e., spectrum splitting), by simply changing the set of interfering BSs $\mathcal{B} \setminus \{b(x, \mathcal{B})\}$ into $\mathcal{B}_M \setminus \{b(x, \mathcal{B})\}$ if $b(x, \mathcal{B}) \in \mathcal{B}_M$ and $\mathcal{B}_m \setminus \{b(x, \mathcal{B})\}$ if $b(x, \mathcal{B}) \in \mathcal{B}_m$, respectively.

where σ^2 is noise power and $b(x, \mathcal{B})$ denotes the index of the BS at location x that provides the highest signal strength, i.e., $b(x, \mathcal{B}) = \arg\max_{b \in \mathcal{B}} E_b(x)$. If there are more than one BSs providing the same highest signal strength, then any suitable tie-breaking rule is used, e.g., choosing the lower indexed BS.

Following Shannon's formula, spectral efficiency at location x is given by:

$$C(x, \mathcal{B}) = \log_2 \left(1 + \Gamma(x, \mathcal{B})\right). \quad \text{[bits/sec/Hz]} \tag{2}$$

2.1.3. Coverage

Let us denote by $\mathcal{A}_{i>j}$ the continuous set of locations that have better SINR from BS i than j. We further denote by $\mathcal{A}_{i=j}$ the set of boundaries having the same SINR from both BSs i and j. Then, the set of locations covered by BS $k \in \mathcal{B}$ (or simply, coverage) can be written as[2]:

$$\begin{aligned} \mathcal{A}_k(\mathcal{B}) &\doteq \{x | x \in \mathcal{A} \text{ s.t. } b(x, \mathcal{B}) = k\} \\ &= \bigcap_{b \in \mathcal{B} \setminus \{k\}} \mathcal{A}_{k>b}. \end{aligned} \tag{3}$$

2.1.4. Area Spectral Efficiency

We adopt the area spectral efficiency (ASE) metric firstly introduced in [27] as our performance metric, which is defined as the summation of the spectral efficiency over the reference area $|\mathcal{A}|$:

$$S_{\mathcal{A}}(\mathcal{B}) \doteq \frac{\sum_{x \in \mathcal{X}} C(x, \mathcal{B}) \cdot Pr(x)}{|\mathcal{A}|}, \quad \text{[bit/sec/Hz/m}^2\text{]} \tag{4}$$

where $Pr(x)$ is the probability of the MT being at a specific location x and $\mathcal{X}$ is the set of user locations included in the area $\mathcal{A}$ satisfying $Pr(x) > 0$ for all $x \in \mathcal{X}$ in the area $\mathcal{A}$. For now, we assume the homogeneous user distribution such that the discrete set $\mathcal{X}$ is a rectangular lattice with a small grid size and the probability of each location is the same. But we will discuss the inhomogeneous case later in Section 3.3.

2.2. General Problem Statement

Let us consider an area of interest $\mathcal{A}$ served by a mobile operator whose access network consists of only macro BSs, say $\mathcal{B}_M$. We assume that the daily traffic profile repeats periodically as a natural effect of users' daily basis[3], and that the required ASE S_{th}^t over time t corresponding to the traffic profile is already known. Suppose that the maximum required ASE $S_{th}^{t^*}$ at the peak time $t^* = \arg\max_t S_{th}^t$ during a day $t \in [t_0, t_0 + D)$ almost approaches to the one that can be provided by turning on all the macro BSs $\mathcal{B}_M$, i.e., $S_{\mathcal{A}}(\mathcal{B}_M) \simeq S_{th}^{t^*}$. Thus, the operator wants to upgrade its access

[2]For simplicity, we ignore the boundaries in the definition of coverage.

[3]In [5, 10, 28], it has been observed that the traffic during the daytime is much higher than that at nighttime. In addition, the traffic profile on a weekend period is much lower than that of a normal weekday. See Fig. 8.

network by deploying additional micro BSs which are considered as the cost-effective way of incrementally increasing capacity inside the initial macro cell deployment.

General problem: We want to minimize the total BS energy consumption during a day while providing $\zeta \geq 1$ times higher ASE than before the upgrade. We can mathematically formulate this problem as the following optimization problem:

$$\begin{aligned} \textbf{(P)} \quad & \min_{\{\mathcal{B}^t\}} \int_{t_0}^{t_0+D} \left(P_M \cdot \left|\mathcal{B}_M^t\right| + P_m \cdot \left|\mathcal{B}_m^t\right|\right) dt \\ & \text{s.t. } S_{\mathcal{A}}(\mathcal{B}^t) \geq \zeta \cdot S_{th}^t, \ \ \forall t \in [t_0, t_0+D), \end{aligned} \tag{5}$$

where $\mathcal{B}^t$ denotes the set of BSs that are turned on at time t; P_M and P_m are the operational power consumption of macro and micro BSs, respectively.

Problem separation: The above general problem **(P)** can be separated into two subproblems: **(P1)** micro BSs *deployment* problem considering the traffic at peak time t^* and **(P2)** BSs *operation* problem during the off-peak period $t \neq t^*$.

It is desirable for the operator to minimize the cost for expanding its infrastructures while guaranteeing the required ASE. Thus, the first problem is to find a minimal deployment of micro BSs which can support the peak time ASE. Note that this deployment issue is an offline problem that can be handled in a centralized network coordinator. Once the micro BS deployment is done, the next problem is how to efficiently operate these micro BSs along with the existing macro BSs for energy conservation during the off-peak period. The solutions for the operation problem should be online distributed algorithms in order to be implemented in practical systems. We will deal with these two problems one by one in the following consecutive sections.

As we will show later in Section 5.2, such a problem separation is not merely for convenience for solving the problem, but also for saving CAPEX and OPEX at the same time from the perspective of mobile operators.

3. Micro BS Deployment Strategy

First, we aim at finding a minimal deployment of micro BSs (i.e., minimizing the total power consumption) while satisfying the raised ASE requirement at the peak time, $t = t^*$:

$$\textbf{(P1)} \quad \min_{\mathcal{B}_m} \ |\mathcal{B}_m| \tag{6}$$

$$\text{s.t.} \quad S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_m) \geq \zeta \cdot S_{th}^{t^*} = \zeta \cdot S_{\mathcal{A}}(\mathcal{B}_M). \tag{7}$$

It is worthwhile mentioning that **(P1)** can be also interpreted as a problem of CAPEX minimization. The deployment problem **(P1)** is basically a combinatorial problem, and that makes it very difficult to find an optimal solution, especially, if the number of candidate locations is large. Therefore, in this paper, our goal is to develop a simple and efficient algorithm.

3.1. Key Observations

We shall start by presenting several observations from various topologies which help us to gain insight and develop an efficient algorithm. To obtain more realistic observations, we acquired the real macro BS topologies in the part of Korea[29] and Manchester, UK [5, 30] as well as typical hexagonal and random topologies. Listed here is brief information about the number of macro BSs and the size of observation area in the topologies that we used: (i) Korea-A: 7 BSs in 5 x 5 km^2, (ii) Korea-B: 15 BSs in 4.5 x 4.5 km^2, (iii) UK: 6 BSs in 2.5 x 2.5 km^2, (iv) hexagonal: 7 BSs in 4 x 4 km^2, and (v) random: 6 BSs in 5 x 5 km^2. It should be mentioned that we ran simulations in a much larger area than the above observation area to avoid edge effects.

We focus on the deployment of *one new micro BS* in the area that is covered by the existing set of macro BSs. The contour plot in Fig. 1 shows how much ASE a micro BS can improve according to the location of deployment. Although this is a snapshot from the topology of Korea-A, similar trends could be observed in the other topologies as well.

Observation 3.1. *As long as a new micro BS is placed not too close to the one of existing BSs that would interfere with each other, ASE can be expected to increase compared to its value before the upgrade. In particular, the ASE increment becomes large as the distances from macro BSs increase.*

The mobile operators are supposed to deploy a micro BS at the location where ASE can be improved. Therefore, throughout the paper, we only consider the set of candidate locations $\mathcal{K}$ for the micro BS deployment as follows:

$$\forall k \in \mathcal{K}, \quad S_{\mathcal{A}}(\mathcal{B} \cup \{k\}) > S_{\mathcal{A}}(\mathcal{B}), \tag{8}$$

Now we examine how much area the micro BS can cover according to the location of deployment and investigate the correlation with ASE increment. In Fig. 2(a), ASE increment has a distinct tendency to increase with coverage. More importantly, it becomes sharper as the coverage increases and this trend can be verified over the other topologies as well in the quantile plots in Fig. 2(b). Note that we plot both cases where macro and micro BSs use the same frequency band as well as they use different frequency bands. This trend is desirable because we are especially interested in the locations that give high performance improvement. In such locations with small variance, we can almost surely assert that coverage and ASE increment have a near-monotonic relationship. Results from monotone test[4] (90.4$\sim$97.0% depending on the topologies) also support the following observation.

Observation 3.2. *When a larger area is covered by a new micro BS, the ASE increment is likely to be higher.*

Motivated by this observation, we assume that the following monotone relationship holds throughout the paper.

$$\begin{aligned} &|\mathcal{A}_k(\mathcal{B} \cup \{k\})| \geq |\mathcal{A}_{k'}(\mathcal{B}' \cup \{k'\})| \quad \Rightarrow \\ &\quad S_{\mathcal{A}}(\mathcal{B} \cup \{k\}) - S_{\mathcal{A}}(\mathcal{B}) \geq S_{\mathcal{A}}(\mathcal{B}' \cup \{k'\}) - S_{\mathcal{A}}(\mathcal{B}'), \end{aligned} \tag{9}$$

[4] We randomly pick two points having positive ASE increments in Fig. 2(a) and check whether the slope between these points are positive or not.

where k (or k') is the index of the micro BS.

These two observations are intuitively understandable. Consider the area covered by the micro BS far from existing macro BSs. Since the signals from the macro BSs are weak, the micro BS will provide the highest SINR to a large extent area. In addition to this large coverage, the area originally had low spectral efficiency, resulting in the high increment of ASE.

Prior to introducing a natural greedy algorithm for **(P1)**, we define a real-valued set function $F : \mathcal{B}_m \to \mathbb{R}$ as follows:

$$F(\mathcal{B}_m) \doteq S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_m) - S_{\mathcal{A}}(\mathcal{B}_M), \tag{10}$$

which returns the ASE increment by incrementally deploying the set of micro BSs $\mathcal{B}_m$.

Definition 3.1. *A real-valued set function H, defined on subsets of a finite set $\mathcal{S}$ is called* submodular *if for all $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \mathcal{S}$ and for all $s \in \mathcal{S} \setminus \mathcal{B}_2$, if it satisfies that*

$$H(\mathcal{B}_1 \cup s) - H(\mathcal{B}_1) \geq H(\mathcal{B}_2 \cup s) - H(\mathcal{B}_2). \tag{11}$$

Submodularity, informally, is an intuitive notion of *diminishing returns*, which states that adding an element to a small set helps more than adding that same element to a larger set. Other equivalent definitions for submodularity can be found in [31].

Lemma 3.2. *The ASE increment function F defined in* (10) *is submodular.*

Proof. In order to prove the submodularity, it is equivalent to check that for all $\mathcal{B}_m \subseteq \mathcal{B}_{m'} \subseteq \mathcal{K}$ and for an arbitrary chosen $k \in \mathcal{K} \setminus \mathcal{B}_{m'}$, the following condition

$$F(\mathcal{B}_m \cup \{k\}) - F(\mathcal{B}_m) \geq F(\mathcal{B}_{m'} \cup \{k\}) - F(\mathcal{B}_{m'}) \tag{12}$$

holds. F is an increasing function by the assumption (8). Hence, we have

$$F(\mathcal{B}_m) \leq F(\mathcal{B}_{m'}). \tag{13}$$

We further have the following inequality because $\mathcal{B}_m$ is the subset of $\mathcal{B}_{m'}$.

$$\begin{aligned} \left|\mathcal{A}_k(\mathcal{B}_M \cup \mathcal{B}_m \cup \{k\})\right| &= \left|\bigcap\nolimits_{b \in \mathcal{B}_M \cup \mathcal{B}_m} \mathcal{A}_{k>b}\right| \\ &\geq \left|\bigcap\nolimits_{b \in \mathcal{B}_M \cup \mathcal{B}_{m'}} \mathcal{A}_{k>b}\right| \\ &= \left|\mathcal{A}_k(\mathcal{B}_M \cup \mathcal{B}_{m'} \cup \{k\})\right|. \end{aligned} \tag{14}$$

By the definition of F and the assumption (9), the coverage inequality (14) can be converted into the ASE inequality:

$$\begin{aligned} F(\mathcal{B}_m \cup \{k\}) &= S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_m \cup \{k\}) - S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_m) \\ &\geq S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_{m'} \cup \{k\}) - S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_{m'}) \\ &= F(\mathcal{B}_{m'} \cup \{k\}). \end{aligned} \tag{15}$$

Combining (13) and (15) completes the proof of submodularity condition (12). □

3.2. Constant-Factor Approximation Greedy Deployment Algorithm

Our proposed greedy deployment algorithm (GDA) starts with the empty set $\mathcal{B}_m^{\text{GDA}} = \emptyset$, and iteratively adds the micro BS location one by one that has the highest increment among the set of candidate locations $\mathcal{K}$ until ASE reaches a target value, i.e., satisfying the constraint (7).

(GDA) Greedy deployment algorithm

1: Initialize $\mathcal{B}_m^{\text{GDA}} = \emptyset$

2: **do while** $S_{\mathcal{A}}(\mathcal{B}_M \cup \mathcal{B}_m^{\text{GDA}}) < \zeta \cdot S_{th}^{t^*}$

3: $\quad k^* = \arg \max_{k \in \mathcal{K} \setminus \mathcal{B}_m^{\text{GDA}}} F(\mathcal{B}_m^{\text{GDA}} \cup \{k\}) - F(\mathcal{B}_m^{\text{GDA}})$,

4: $\quad \mathcal{B}_m^{\text{GDA}} \leftarrow \mathcal{B}_m^{\text{GDA}} \cup \{k\}$

5: **end do**

Theorem 3.1. *The ASE increment achieved by an optimal deployment with the same number of micro BSs as the greedy algorithm cannot be more than a factor of* $e/(e-1)$ *from the ASE increment achieved by the greedy algorithm.*

$$\max_{|\mathcal{B}_m| = |\mathcal{B}_m^{GDA}|} F(\mathcal{B}_m) \leq \frac{e}{e-1} F(\mathcal{B}_m^{GDA}), \tag{16}$$

where the constant e *is base of the natural logarithm.*

Proof. Let $\mathcal{B}_m^{\text{GDA}} = \{b_1, \ldots, b_{|\mathcal{B}_m^{\text{GDA}}|}\}$ and $\mathcal{B}_m^* = \{b_1^*, \ldots, b_{|\mathcal{B}_m^{\text{GDA}}|}^*\}$ denote greedy and optimal solutions, respectively. We define $\mathcal{B}_{m,i}^{\text{GDA}} = \{b_1, \ldots, b_i\}$ and $\mathcal{B}_{m,i}^* = \{b_1^*, \ldots, b_i^*\}$ for $i = 1, \ldots, |\mathcal{B}_m^{\text{GDA}}|$. And further define $\mathcal{B}_{m,0}^{\text{GDA}} = \emptyset$ and $\mathcal{B}_{m,0}^* = \emptyset$. Then, for all $i = 0, \ldots, |\mathcal{B}_m^{\text{GDA}}|$, we can obtain eq. (17).

In (17), the first inequality is due to the increasing property of F, the second equality is a simple telescoping sum, and the third equality is trivial by definition. The fourth inequality is a direct application of submodularity of F from Lemma 3.2, with $\mathcal{B}_{m,i}^{\text{GDA}} \subseteq \mathcal{B}_{m,i}^{\text{GDA}} \cup \mathcal{B}_{m,j-1}^*$, and the fifth inequality is direct from the definition of greedy algorithm.

Now define $\Delta_i \doteq F(\mathcal{B}_m^*) - F(\mathcal{B}_{m,i}^{\text{GDA}})$, then (17) can be rewritten by $\Delta_i \leq |\mathcal{B}_m^{\text{GDA}}| \cdot (\Delta_i - \Delta_{i+1})$, or equivalently,

$$\Delta_{i+1} \leq \left(1 - 1/|\mathcal{B}_m^{\text{GDA}}|\right) \Delta_i, \tag{18}$$

for all $i = 0, \ldots, |\mathcal{B}_m^{\text{GDA}}|$. Hence,

$$\Delta_{|\mathcal{B}_m^{\text{GDA}}|} \leq \left(1 - 1/|\mathcal{B}_m^{\text{GDA}}|\right)^{|\mathcal{B}_m^{\text{GDA}}|} F(\mathcal{B}_m^*) \tag{19}$$

$$\leq \frac{1}{e} F(\mathcal{B}_m^*). \tag{20}$$

By putting $\Delta_{|\mathcal{B}_m^{\mathrm{GDA}}|} = F(\mathcal{B}_m^*) - F(\mathcal{B}_m^{\mathrm{GDA}})$ into (20), we have the following:

$$F(\mathcal{B}_m^*) \leq \frac{e}{e-1} F(\mathcal{B}_m^{\mathrm{GDA}}). \tag{21}$$

This completes the proof. □

So far we have assumed that micro BSs have the same operational power P_m. However, the above constant-factor approximation result can be extended to general cases [32] with *different* powers (say, $P_k = P_i$ for $k \in \mathcal{B}_i$) as follows.

Corollary 3.1. *For a general case where various types of BSs such as macro, micro, pico and even femto BSs having different operational powers coexist in a complex manner. The GDA only needs to be modified as follows:*

$$k^* = \arg \max_{k \in \mathcal{K} \setminus \mathcal{B}_m^{GDA}} \frac{F(\mathcal{B}_m^{GDA} \cup \{k\}) - F(\mathcal{B}_m^{GDA})}{P_k}. \tag{22}$$

Note that this metric can be interpreted as finding the location with the *highest ASE increment per unit power consumption*.

3.3. Inhomogeneous Traffic Case

Submodularity: The submodularity of ASE increment function under homogeneous traffic distributions allow us to derive Theorem 3.1. Even though the theorem does not hold in the inhomogeneous case anymore where $Pr(x)$ is not the same over the area, we present some numerical results instead. For the test, an inhomogeneous scenario having five randomly generated hot-spots ($100 \times 100\mathrm{m}^2$) in the Korea-A topology is considered. In this inhomogeneous scenario, we investigate the probability that the submodularity condition $F(\mathcal{B}_m \cup \{k\}) - F(\mathcal{B}_m) \geq F(\mathcal{B}_{m'} \cup \{k\}) - F(\mathcal{B}_{m'})$ for all $\mathcal{B}_m \subseteq \mathcal{B}_{m'} \subseteq \mathcal{K}$ and for an arbitrary chosen $k \in \mathcal{K} \setminus \mathcal{B}_{m'}$ holds.

Fig. 3 shows the CDF of the difference in ASE increment and the inner figure shows the scatter plot for details. It is clear that we lose the submodularity because

---

$$\begin{aligned}
F(\mathcal{B}_m^*) \;&\leq\; F(\mathcal{B}_{m,i}^{\mathrm{GDA}} \cup \mathcal{B}_m^*) \;=\; F(\mathcal{B}_{m,i}^{\mathrm{GDA}}) + \sum_{j=1}^{|\mathcal{B}_m^{\mathrm{GDA}}|} \left[F(\mathcal{B}_{m,i}^{\mathrm{GDA}} \cup \mathcal{B}_{m,j}^*) - F(\mathcal{B}_{m,i}^{\mathrm{GDA}} \cup \mathcal{B}_{m,j-1}^*)\right] \\
&=\; F(\mathcal{B}_{m,i}^{\mathrm{GDA}}) + \sum_{j=1}^{|\mathcal{B}_m^{\mathrm{GDA}}|} \left[F((\mathcal{B}_{m,i}^{\mathrm{GDA}} \cup \mathcal{B}_{m,j-1}^*) \cup \{b_j^*\}) - F(\mathcal{B}_{m,i}^{\mathrm{GDA}} \cup \mathcal{B}_{m,j-1}^*)\right] \\
&\leq\; F(\mathcal{B}_{m,i}^{\mathrm{GDA}}) + \sum_{j=1}^{|\mathcal{B}_m^{\mathrm{GDA}}|} \left[F(\mathcal{B}_{m,i}^{\mathrm{GDA}} \cup \{b_j^*\}) - F(\mathcal{B}_{m,i}^{\mathrm{GDA}})\right] \\
&\leq\; F(\mathcal{B}_{m,i}^{\mathrm{GDA}}) + |\mathcal{B}_m^{\mathrm{GDA}}| \cdot \left[F(\mathcal{B}_{m,i+1}^{\mathrm{GDA}}) - F(\mathcal{B}_{m,i}^{\mathrm{GDA}})\right].
\end{aligned} \tag{17}$$

about 19% of locations violates the inequality. However, ASE decrements in such cases are relatively small and most of locations ($> 80\%$) still satisfy the condition. Thus, we conjecture that GDA still works well under inhomogeneous traffic distributions, and later in Section 5.2, we provide an affirmative simulation result supporting this.

Hot spots or in-building scenarios: **The way how GDA works is sequentially finding the location with highest ASE increment per unit power consumption. When calculating ASE (see eq.** (4) **for its definition), we can naturally capture the hot spot and in-building scenarios. For example, in hot spot areas,** $Pr(x)$ **is likely to be higher, so is its impact on the overall ASE. Thus, GDA tends to deploy new micro BSs in or near the hot spot areas. Similarly, it is likely that GDA will choose the location in or near the building areas. Users in the building suffer from lower spectral efficiency** $\mathcal{C}(x, B)$ **due to wall penetration loss (usually** > 5**dB) compared to outdoor users having the similar distance to the serving BS. Thus, deploying new micro BSs nearby would bring large ASE increment.**

4. Dynamic BS Operation Strategy

Since BSs are typically provisioned to handle the peak time traffic, they will be under-utilized at most of off-peak times, $t \neq t^*$. In other words, a considerable amount of energy will be wasted unless an appropriate dynamic BS operation algorithm is not employed.

Thus, our objective at the off-peak period is to find a dynamic operation of BSs that minimizes the total operational power consumption (i.e., OPEX minimization) while satisfying the raised ASE requirement [5]

$$\begin{aligned} \textbf{(P2)} \quad & \min_{\mathcal{B}^t} P_M \cdot \left|\mathcal{B}_M^t\right| + P_m \cdot \left|\mathcal{B}_m^t\right| \\ & \text{s.t.} \;\; S_{\mathcal{A}}(\mathcal{B}^t) \geq \zeta \cdot S_{th}^t. \end{aligned} \tag{23}$$

The operation problem **(P2)** is a combinatorial optimization problem as well. Thus, in the following consecutive subsection, we propose a suboptimal offline centralized algorithm and two online distributed BS switching algorithms.

4.1. Centralized BS Switching Algorithm

Due to the similarity between the deployment and operation problems in nature, we may use the generalized deployment algorithm in (22) as a centralized offline switching algorithm for the operation problem as follows:

$$k^* = \arg\max_{k \in \mathcal{B}^t} \frac{F(B^t \cup \{k\}) - F(B^t)}{P_k}. \tag{24}$$

This centralized algorithm has a nice feature of the constant-factor approximation; however, it not only requires a lot of feedbacks from all BSs to the network coordinator but also should be restarted from the empty set (i.e., turning off all BSs), which

[5] Since the traffic profile usually remains stationary for quite a long time [28], the original problem may be solved whenever there is a big change in traffic, e.g, every hour or even less frequently.

makes it difficult for the centralized algorithm to be implemented in practice. In order to overcome such difficulties, we consider the design of distributed online algorithms.

4.2. Distributed BS Switching Algorithms

Using the Lagrangian relaxation with a multiplier λ, the BS operation problem **(P2)** can be separated by the summation of the switching problem at each BS as follows.

$$\begin{aligned} L(\mathcal{B}^t, \lambda) &= \sum_{b \in \mathcal{B}^t} P_b + \lambda \left[\zeta \cdot S_{th}^t - S_{\mathcal{A}}(\mathcal{B}^t) \right] \\ &= \sum_{b \in \mathcal{B}^t} \underbrace{\left[P_b a_b^t + \frac{\lambda}{|\mathcal{A}|} \Big(\frac{\zeta \cdot |\mathcal{A}|}{|\mathcal{B}|} S_{th}^t - \sum_{x \in \mathcal{X}_b} C(x, \mathcal{B}^t) \Big) \right]}_{L_b(a_b^t, \lambda)}, \end{aligned}$$

where a_b^t denotes the indicator of BS status, i.e., $a_b^t = 1$ when the BS b is on at time t, and 0 otherwise; $\mathcal{X}_b$ denotes the set of locations included in the serving area of BS b, and

$$L_b(a_b^t, \lambda) = \begin{cases} \frac{\lambda}{|\mathcal{A}|} \Big[\frac{\zeta \cdot |\mathcal{A}|}{|\mathcal{B}|} S_{th}^t - \sum_{x \in \mathcal{X}_b} C(x, \mathcal{B}^t) \Big] + P_b, & \text{if } a_b^t = 1. \quad \text{(25a)} \\ \frac{\lambda}{|\mathcal{A}|} \Big[\frac{\zeta \cdot |\mathcal{A}|}{|\mathcal{B}|} S_{th}^t - \sum_{x \in \mathcal{X}_b} C(x, \mathcal{B}^t - \{b\}) \Big], & \text{otherwise.} \quad \text{(25b)} \end{cases}$$

To minimize the relaxation gap, the network coordinator updates the Lagrangian multiplier λ based on gradient descent iterative method with a small step size $\epsilon > 0$, i.e.,

$$\lambda \leftarrow \lambda + \epsilon \left[\zeta \cdot S_{th}^t - S_{\mathcal{A}}(\mathcal{B}^t) \right], \tag{26}$$

where $S_{\mathcal{A}}(\mathcal{B}^t)$ can be calculated by collecting the local ASE from each BS as follows:

$$S_{\mathcal{A}}(\mathcal{B}^t) = \frac{1}{|\mathcal{A}|} \sum_{b \in \mathcal{B}^t} |\mathcal{A}_b| \cdot S_{\mathcal{A}_b}(\mathcal{B}^t). \tag{27}$$

For any given λ, BS b needs to be turned off for energy saving if the difference between (25a) and (25b) is less than or equal to zero, i.e.,

$$\begin{aligned} & L_b(0, \lambda) - L_b(1, \lambda) \leq 0 \\ \Leftrightarrow \; & \frac{|\mathcal{A}_b| \cdot \{ S_{\mathcal{A}_b}(\mathcal{B}^t) - S_{\mathcal{A}_b}(\mathcal{B}^t - \{b\}) \}}{P_b} \leq \frac{|\mathcal{A}|}{\lambda}. \end{aligned} \tag{28}$$

This condition (28) can be interpreted as follows: (i) The less decrement in spectral efficiency (i.e., small impact on QoS) the BS has and/or (ii) the larger operational power (i.e., large energy saving) the BS consumes, the more likely the BS is switched off.

With help of the switching off condition in (28), we propose a distributed BS switching algorithm (S-OFF1). As each BS locally determines its own on-off state,

this algorithm reduces signaling overhead compared to the centralized algorithm given in (24) requiring high message passing bandwidth to a centralized node where the on-off decision is made.

(S-OFF1) SINR-based distributed switching algorithm

At each time t, each BS b reports current local information $|\mathcal{A}_b| \cdot S_{\mathcal{A}_b}(\mathcal{B}^t)$ to the network coordinator, and receives the Lagrangian multiplier λ. If the performance decrement in spectral efficiency per unit operational power is less than a certain threshold, then the BS b will be switched off.

$$\frac{|\mathcal{A}_b| \cdot \{S_{\mathcal{A}_b}(\mathcal{B}^t) - S_{\mathcal{A}_b}(\mathcal{B}^t - \{b\})\}}{P_b} \leq \frac{|\mathcal{A}|}{\lambda}. \tag{29}$$

The BS switching-on procedure can be accomplished by the reverse way of the switching-off procedure. Without any additional calculation in off-state, the BS b is switched on when the target ASE reaches the same value that the BS was originally switched off.

(S-OFF1) requires SINR estimations from MTs in its coverage before and after turning off BS b when calculating $S_{\mathcal{A}_b}(\mathcal{B}^t)$ and $S_{\mathcal{A}_b}(\mathcal{B}^t - \{b\})$, respectively. We further propose (S-OFF2) that is based on SNR estimation. Its computation on MTs is simpler than (S-OFF1) because it does not require the total interference but only requires the best signal strength, the second best signal strength and noise. Besides, in a real cellular system, since MTs measure the best and second signal strengths for their mobility management (e.g., handover), all required measurements for (S-OFF2) are readily available. Thus, we claim that (S-OFF2) is practical despite slight loss in energy-efficiency due to its conservative operation (See Proposition 4.1).

(S-OFF2) SNR-based distributed switching algorithm

$$\frac{|\mathcal{A}_b| \cdot \left\{S^{\sigma^2}_{\mathcal{A}_b}(\mathcal{B}^t) - S^{\sigma^2}_{\mathcal{A}_b}(\mathcal{B}^t - \{b\})\right\}}{P_b} \leq \frac{|\mathcal{A}|}{\lambda}. \tag{30}$$

where $S^{\sigma^2}_{\mathcal{A}}(\mathcal{B}) = \frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{X}} \log_2(1 + E_{b(x,\mathcal{B})}(x)/\sigma^2)$.

Proposition 4.1. *The ASE differences of SINR-based and SNR-based distributed algorithms satisfy the following:*

$$S_{\mathcal{A}_b}(\mathcal{B}^t) - S_{\mathcal{A}_b}(\mathcal{B}^t - \{b\}) \leq S^{\sigma^2}_{\mathcal{A}_b}(\mathcal{B}^t) - S^{\sigma^2}_{\mathcal{A}_b}(\mathcal{B}^t - \{b\}). \tag{31}$$

Accordingly, (S-OFF1) *can turn off more BSs than or equal to* (S-OFF2)*, which results in better energy-efficiency.*

Proof. Denote the indexes of BSs that provides the best and the second best signal strength by $k_1 = b(x, \mathcal{B}^t)$ and $k_2 = b(x, \mathcal{B}^t - \{k_1\})$, respectively. And further denote the total amount of interference at location x from all BSs except BSs k_1 and k_2 by $I_x = \sum_{b\in\mathcal{B}\setminus\{k_1,k_2\}} E_b(x)$. Then, the difference in ASE of two algorithms at location

x can be expressed as:

$$
\begin{aligned}
& C(x, \mathcal{B}^t) - C(x, \mathcal{B}^t - \{b\}) \\
&= \log_2 \left\{ 1 + \frac{E_{k_1}(x)}{E_{k_2}(x) + I_x + \sigma^2} \right\} - \log_2 \left\{ 1 + \frac{E_{k_2}(x)}{I_x + \sigma^2} \right\} \\
&\leq \log_2 \left\{ 1 + \frac{E_{k_1}(x)}{I_x + \sigma^2} \right\} - \log_2 \left\{ 1 + \frac{E_{k_2}(x)}{I_x + \sigma^2} \right\} \\
&= \log_2 \left\{ \frac{I_x + \sigma^2 + E_{k_1}(x)}{I_x + \sigma^2 + E_{k_2}(x)} \right\} \\
&\leq \log_2 \left\{ \frac{\sigma^2 + E_{k_1}(x)}{\sigma^2 + E_{k_2}(x)} \right\} \\
&= C^{\sigma^2}(x, \mathcal{B}^t) - C^{\sigma^2}(x, \mathcal{B}^t - \{b\}) \qquad (32)
\end{aligned}
$$

The first inequality is due to $E_{k_2}(x) > 0$ and the last inequality holds because $\log_2 \left(\frac{z+E_{k_1}(x)}{z+E_{k_2}(x)} \right)$ is a monotone decreasing convex function of z when $E_{k_1}(x) > E_{k_2}(x)$ [33]. By substituting the definition of ASE into (32), we can obatin the condition (31). □

5. Numerical Results

For our simulation, we consider a topology with 17 macro BSs in $3.5 \times 3.5\text{km}^2$ as shown in Fig. 5. It is a part of actual network infrastructure operated by one of the major mobile operators in UK [5, 30]. To avoid edge effects, observations are made only in the center area of $2.5 \times 2.5\text{km}^2$, which is referred as $\mathcal{A}$. MTs are assumed to be uniformly distributed in the observation area, i.e., $Pr(x) = 1,\ x \in \mathcal{A}$, but the case of inhomogeneous traffic distribution is also considered in Fig. 6. In modeling the propagation environment, the modified COST 231 Hata path loss model with macro BS height $h = 32$m and micro BS height $h = 12.5$m is used. The other parameters for the simulations follow the suggestions in the IEEE 802.16m evaluation methodology document [34].

5.1. Base Station Deployment

We first investigate the performance improvement according to the additional deployment by the proposed greedy deployment algorithm (GDA) on top of the existing deployment of macro BSs. Four types of BSs with the typical values of transmission power [35, 36] are considered: the macro BS with transmission power p_M of 43dBm and the micro BSs with transmission power p_m of 33dBm, 30dBm and 27dBm, respectively. All intersection points on a grid with 50m are considered as a set of candidate locations for the deployment.

As can be clearly seen in Fig. 4, there are diminishing returns on the normalized ASE increment. This is not only because the coverage of newly deployed BS will shrink but the amount of interference in the network also increases as the number of BSs grows. To meet the target ASE increment of 10%, while only five additional macro BSs are needed, 15, 22 or 30 micro BSs (three to six times more than the case

of macro BSs) are needed depending on their transmission powers. Nevertheless, the transmission power consumptions (p_M and p_m) of additional micro BSs are much less than that of additional macro BSs.

For example, while a total of 100W (=43dBm $\times$ 5 macro BSs) is consumed by the macro BSs, only 30W, 22W, or 15W is consumed by the each type of micro BSs. When we reflect the total operational power consumptions[6], the advantage of energy-efficient micro BSs becomes more clear. Table 1 shows the additional total power consumptions for different target ASE increments. Compared to the case of macro BSs, deploying micro BSs can reduce more than 3kW and 7kW for the target ASE increment 10% and 15%, respectively.

5.2. Justification for Problem Separation into (P1) and (P2)

Throughout the paper, instead of directly solving the original problem **(P)**, we have separated it into two subproblems and tackled them step by step. In this section, we will compare our two-step approach (GDA + S-OFF1/S-OFF2) with the optimal approach in terms of the total energy consumption and other performance metrics to justify the separation.

To this end, we first consider an optimal exhaustive search (OES) that provides the highest ASE increment for a given number of additional micro BS. Let us denote its solution (i.e., the set of k additional micro BSs) by $\mathcal{B}_m(k)$ and further denote the resulting ASE by ASE_k. If we find the minimal number of k satisfying the following ASE requirement at each time t,

$$\text{ASE}_{k-1} < \zeta \cdot S_{th}^t \leq \text{ASE}_k$$

and make the set of micro BSs $\mathcal{B}_m(k)$ to be active, then it will consume the least additional power consumption. In other words, this is an optimal solution of **(P)**.

However, it is virtually impossible to implement the OES in practice due to its prohibitive complexity. For example, let us denote the cardinality of the set of all candidate deployment locations for micro BSs by $|\mathcal{K}|$. There are $\binom{|\mathcal{K}|}{i}$ possible different ways to choose i micro BSs among $|\mathcal{K}|$ locations. The total complexity of OES becomes $\sum_{i=1}^{k}\binom{|\mathcal{K}|}{i}$ from 1 to k micro BSs deployment, which grows exponentially with the cardinality, i.e., $O(2^{|\mathcal{K}|})$. Note that the GDA only requires polynomial time complexity $\sum_{i=1}^{k}(|\mathcal{K}|-i+1) \sim O(|\mathcal{K}|^2)$ for the deployment. When deploying up to $k=6$ micro BSs out of $|\mathcal{K}|=2500$ positions (a grid with 50m in $2.5\times 2.5\text{km}^2$), the complexity of each algorithm will be: $\sum_{i=1}^{5}\binom{2500}{i} = 3.4\times 10^{17}$ for OES and $\sum_{i=1}^{5}(2501-i) = 14985$ for GDA.

It should be mentioned that despite such a low complexity, GDA performs very close to the optimum. In order to make a comparison, micro BSs are restricted to be deployed among $|\mathcal{K}|$ candidate locations (up to 100, instead of the grid with 50m) that

[6]Based on the relationship between transmission and operational power consumptions given in [35, 36], we calculate the total operational powers for all types of BSs. For example, P_M = 865W for the macro BSs with the transmission power of 43dBm; P_m = 35W, 38W and 43W for the micro BSs with the transmission powers of 27dBm, 30dBm and 33dBm, respectively.

are randomly generated in the observation area. Fig. 6 shows the gap of the ASE increment between the OES and GDA, defined by $\frac{F(\mathcal{B}_m^{\text{OES}})-F(\mathcal{B}_m^{\text{GDA}})}{F(\mathcal{B}_m^{\text{OES}})} \times 100$ [%], after the deployment of eight[7] micro BSs. Even though the gap increases as the number of candidate locations increases, its absolute value is not only small (e.g., less than 0.011% in the homogenous traffic case) but also its rate of change is decreasing (i.e., concave). Moreover, based on our theoretical result in Theorem 3.1, we know it should be upper bounded by $\frac{F(\mathcal{B}_m^{\text{OES}})-F(\mathcal{B}_m^{\text{GDA}})}{F(\mathcal{B}_m^{\text{OES}})} \leq \frac{1}{e} \simeq 0.37$.

We also consider the case of inhomogeneous traffic, in which Theorem 3.1 is no longer valid, to empirically show that GDA still works well. The MT probability $Pr(x)$ is scaled based on distance d from the upper left corner, e.g., $[1 - Pr(x)] \propto d$. In other words, a linearly decreasing traffic load along the diagonal direction from the upper left corner to the lower right corner is generated. Although the gap is slightly higher that of homogeneous traffic case, the overall trend is similar and its absolute value is still very small (e.g., less than 0.016% at $|\mathcal{K}| = 100$). This implies that the proposed GDA is likely to work well in practice, where the assumption of homogeneous traffic does not usually hold.

Last but not least, there is another reason why it makes more sense to employ the two-step approach in practice. TABLE 2 gives a good example how micro BSs are deployed by each algorithm at $|\mathcal{K}| = 100$ (see a pictorial snapshot of Fig. 5 for the locations of deployed micro BSs). GDA incrementally deploys one more BS based on the previous step in an accumulated manner, whereas OES does not. It always finds the best deployment in each step regardless of what have been used so far, and thus ends up deploying more micro BSs than GDA: 6 vs. 7 micro BSs. TABLE 3 shows the average number of micro BSs deployed by each algorithm. As expected, more additional BSs are deployed by OES (i.e., higher CAPEX), especially, when $|\mathcal{K}|$ is large. On the other hand, due to the slightly higher ASE as shown in Fig. 6, OES apparently consumes less energy than GDA in conjunction with S-OFF1. However, the gain of OES over the two-step approach turns out to be very marginal, e.g., $< 0.1 \sim 1\%$ energy reduction based on our simulations at $|\mathcal{K}| = 100$. To sum up, spending 8.3% more CAPEX for micro BSs is not a good idea to obtain a less than 1% energy reduction. In reality, moreover, in addition to one time CAPEX (cost for micro BSs), there is hidden OPEX (monthly site lease, transport, etc.), which is proportional to the number of micro BSs. So, we believe that the two-step approach is a more viable solution.

5.3. Base Station Operation

Now we examine the performance of the proposed BS switching algorithms. In Fig. 7, we compare the performance of the proposed S-OFF algorithms with that of centralized algorithm by the normalized required ASE $S_{th}^t / {S_{th}}^{t^*}$. **To better demonstrate the performance of the proposed algorithms, we also consider SWES algorithm in [7], switching off the BS with the least *network impact* F_b in an iterative manner.**

[7] Due to the computational complexity, the number of micro BS deployment is limited to six only when $|\mathcal{K}| = 100$.

First of all, it is worthwhile mentioning that such simple distributed S-OFF algorithms can closely approximate the complex centralized algorithm. When the normalized required ASE is less than 0.4, the centralized and two distributed S-OFF algorithms consume the same amount of energy. As the normalized required ASE increases over 0.4, performance gaps begin to arise, but are marginal. **For example, at $S_{th}^{t}/{S_{th}}^{t^*}$ = 0.7, additional 1.6% and 2.2% powers are used in two distributed S-OFF algorithms, respectively, while SWES algorithm consumes additional 7.7% over the centralized algorithm. Note also that (S-OFF1) performs better than (S-OFF2) in terms of energy savings, which coincides with Proposition 4.1.**

The superiority of S-OFF over SWES in HetNet environments comes from the fact that S-OFF considers not only the change of traffic load but also the characteristics of HetNet environment from the design stage, e.g., capturing different the total operational power between macro and micro BSs. However, on the other hand, SWES adopts the metric of network impact for its switching decision, which only considers the traffic load increment of each BS. This makes SWES tend to turn off micro BSs (relatively carrying less traffic load) earlier than macro BSs irrespective of their relative load and power consumption to macro BSs.

5.4. Overall Energy Savings Under Real Cellular Temporal Traffic Traces

To obtain more realistic results, we further consider traffic profiles in a metropolitan urban area during one week as shown in Fig. 8 that is recorded by an anonymous mobile operator [5]. The low traffic period (less than 0.4 of the maximum value) is about 65% of time assuming two weekend days in a typical week. This implies that our distributed algorithms can obtain the same performance as the centralized near optimal solution during 65% of time and can still be within about 2% from the centralized near optimal solution during the rest of time.

Note that we used the BS topology for this simulation which is obtained from the proposed deployment algorithm with the target increment of 15% in the previous subsection 5.1. Table 4 summarizes the total energy savings for different algorithms during one day compared to the conventional static operation (i.e., always-ON strategy). It is expected that about 60% and 80% of energy consumption can be reduced by dynamic BS operation on weekday and weekend, respectively. Given that OPEX of wireless network operators for electricity is more than 10 billion dollars globally [5], this could translate to huge economic benefit to the operators.

6. Conclusion

Heterogeneous cellular networks, consisting of cells with different sizes and ranging from macro to micro cells, will play a pivotal role in next-generation wireless networks. It can increase spectral efficiency in a cost-effective and power-efficient manner. In this paper, we have proposed an energy-aware heterogeneous cell deployment and operation framework that has theoretical results as well as practical guidelines on how mobile operators manage their BSs. We have specifically focused on a problem pertaining to total energy consumption minimization while satisfying the requirement of ASE, and decomposed it into deployment problem at peak time and operation

problem at off-peak time. For the deployment problem, we have proposed a constant-factor approximation greedy algorithm. For the operation problem, we have proposed two distributed online switching algorithms using Lagrangian relaxation to have more practical solutions. Extensive simulations based on the acquired real BS topologies and traffic profiles show that the proposed deployment and operation algorithms can dramatically reduce the total energy consumption.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A1A1009526).

References

[1] K. Son, E. Oh, B. Krishnamachari, Energy-aware hierarchical cell configuration: from deployment to operation, in: Proc. IEEE INFOCOM GCN Workshop, Shanghai, China, 2011, pp. 289–294.

[2] J. G. Andrews, The seven ways HetNets are a paradigm shift, IEEE Commun. Mag. 51 (3) (2013) 136–144.

[3] A. Ghosh, J. G. Andrews, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, P. Xia, H. S. Jo, H. S. Dhillon, T. D. Novlan, Heterogeneous cellular networks: From theory to practice, IEEE Commun. Mag. 50 (6) (2012) 54–64.

[4] X. Wu, B. Murherjee, D. Ghosal, Hierarchical architectures in the third-generation cellular network, IEEE Wireless Commun. Mag. 11 (3) (2004) 62–71.

[5] E. Oh, B. Krishnamachari, X. Liu, Z. Niu, Towards dynamic energy-efficient operation of cellular network infrastructure, IEEE Commun. Mag. 49 (6) (2011) 56–61.

[6] M. A. Marsan, L. Chiaraviglio, D. Ciullo, M. Meo, Optimal energy savings in cellular access networks, in: Proc. of IEEE GreenComm, Dresden, Germany, 2009.

[7] E. Oh, K. Son, B. Krishnamachari, Dynamic base station switching-on/off strategies for green cellular networks, IEEE Trans. Wireless Commun. 12 (5) (2013) 2126–2136.

[8] L. Chiaraviglio, D. Ciullo, M. Meo, M. A. Marsan, I. Torino, Energy-aware UMTS access networks, in: Proc. of WPMC Symposium, Lapland, Finland, 2008.

[9] K. Son, H. Kim, Y. Yi, B. Krishnamachari, Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks, IEEE J. Sel. Areas Commun. 29 (8) (2011) 1525–1536.

[10] M. A. Marsan, M. Meo, Energy efficient management of two cellular access networks, in: Proc. of ACM GreenMetrics, Seattle, WA, 2009.

[11] C. Peng, S.-B. Lee, S. Lu, H. Luo, H. Li, Traffic-driven power saving in operational 3G cellular networks, in: Proc. ACM MOBICOM, 2011, pp. 121–132.

[12] K. Son, S. Nagaraj, M. Sarkar, S. Dey, Qos-aware dynamic cell reconfiguration for energy conservation in cellular networks, in: Proc. IEEE WCNC, Shanghai, China, 2013, pp. 2022–2027.

[13] M. Ismail, W. Zhuang, A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium, IEEE J. Sel. Areas Commun. 30 (2) (2012) 425–432.

[14] D. Fooladivanda, C. Rosenberg, Joint resource allocation and user association for heterogeneous wireless cellular networks, IEEE Trans. Wireless Commun. 12 (1) (2013) 248–257.

[15] R. Xie, F. R. Yu, H. Ji, Interference management and power allocation for energy-efficient cognitive femtocell networks, Mobile Networks and Appl. 18 (4) (2013) 578–590.

[16] S. Shen, T. M. Lok, Dynamic power allocation for downlink interference management in a two-tier OFDMA network, IEEE Trans. Veh. Technol. 62 (8) (2013) 4120–4125.

[17] J. Kwak, K. Son, Y. Yi, S. Chong, Greening effect of spatio-temporal power sharing policies in cellular networks with energy constraints, IEEE Trans. Wireless Commun. 11 (12) (2012) 4405–4415.

[18] H. Claussen, Co-channel operation of macro- and femtocells in a hierarchical cell structure, Int. J. Wireless Inf. Networks 15 (3) (2008) 137–147.

[19] R. Mahapatra, A. De Domenico, R. Gupta, E. Calvanese Strinati, Green framework for future heterogeneous wireless networks, Computer Net. 57 (6) (2013) 1518–1528.

[20] D. Stamatelos, A. Ephremides, Spectral efficiency and optimal base placement for indoor wireless networks, IEEE J. Sel. Areas Commun. 14 (4) (1996) 651–661.

[21] A. Srinivas, E. Modiano, Joint node placement and assignment for throughput optimization in mobile backbone networks, in: Proc. IEEE INFOCOM, Phoenix, AZ, 2008.

[22] S.-R. Cho, W. Choi, Energy-efficient repulsive cell activation for heterogeneous cellular networks, IEEE J. Sel. Areas Commun. 31 (5) (2013) 870–882.

[23] Y. S. Soh, T. Q. S. Quek, M. Kountouris, H. Shin, Energy-efficient heterogeneous cellular networks, IEEE J. Sel. Areas Commun. 31 (5) (2013) 840–850.

[24] X. Zhang, Z. Su, Z. Yan, W. Wang, Energy-efficiency study for two-tier heterogeneous networks (HetNet) under coverage performance constraints, Mobile Networks and Appl. 18 (4) (2013) 567–577.

[25] W.-Y. Shin, H. Yi, V. Tarokh, Energy-efficient base-station topologies for green cellular networks, in: Proc. of IEEE CCNC, Las Vegas, NV, 2013, pp. 91–96.

[26] K. Son, B. Krishnamachari, Speedbalance: Speed-scaling-aware optimal load balancing for green cellular networks, in: Proc. IEEE INFOCOM, Orlando, FL, USA, 2012, pp. 2816–2820.

[27] M. S. Alouini, A. J. Goldsmith, Area spectral efficiency of cellular mobile radio systems, IEEE Trans. Veh. Technol. 48 (4) (1999) 1047–1066.

[28] D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz, Primary users in cellular networks: A large-scale measurement study, in: Proc. of IEEE DySPAN, Chicago, IL, 2008.

[29] K. Son, S. Lee, Y. Yi, S. Chong, Practical dynamic interference management in multi-carrier multi-cell wireless networks: A reference user based approach, in: Proc. WiOpt, Avignon, France, 2010.

[30] Sitefinder: Mobile phone base station database.
URL `http://www.sitefinder.ofcom.org.uk/`.

[31] G. Nemhauser, L. Wolsey, M. Fisher, An analysis of the approximations for maximizing submodular set functions-I, Mathematical Programming 14 (1) (1978) 265–294.

[32] M. Sviridenko, A note on maximizing a submodular set function subject to knapsack constraint, Operations Research Letters 32 (2004) 41–43.

[33] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[34] IEEE 802.16m-08/004r5: IEEE 802.16m evaluation methodology document (EMD) (2009).

[35] A. J. Fehske, F. Richter, G. P. Fettweis, Energy efficiency improvements through micro sites in cellular mobile radio networks, in: Proc. of IEEE GreenComm, Honolulu, HI, 2009.

[36] O. Arnold, F. Richter, G. Fettweis, O. Blume, Power consumption modeling of different base station types in heterogeneous cellular networks, in: Proc. of ICT MobileSummit, Florence, Italy, 2010.

Table 1: Additional total power consumptions required for the types of BSs to meet the target ASE increment.

	Macro 43dBm	Micro 33dBm	Micro 30dBm	Micro 27dBm
$\zeta = 1.10$	4325W	645W	836W	1050W
$\zeta = 1.15$	9515W	1476W	1672W	2240W

Table 2: An example of micro BS deployment. Note that micro BS 96 is no longer selected and replaced with micro BS 21 after the 3rd step in the OES.

Deployment step	GDA	OES
1st step	{61}	{61}
2nd step	{61,96}	{61,96}
3rd step	{1,61,96}	{1,61,96}
4th step	{1,61,66,96}	{1,21,61,66}
5th step	{1,49,61,66,96}	{1,21,49,61,66}
6th step	{1,49,61,66,89,96}	{1,21,49,61,66,79}
Deployed micro BSs	{1,49,61,66,89,96}	{1,21,49,61,66,79,96}

Table 3: The number of micro BSs deployed on average.

Candidate locations $\lvert\mathcal{K}\rvert$	20	30	50	100
GDA	8	8	8	6
OES	8.1	8.2	8.4	6.5
Additional micro BSs required	1.3%	2.5%	5.0%	8.3%

Table 4: Energy savings during one day when ζ=1.15.

		Micro 33dBm	Micro 30dBm	Micro 27dBm
Weekday	Cent.	67.5%	67.2%	60.6%
	S-OFF1	67.1%	65.6%	56.2%
	S-OFF2	65.6%	65.4%	55.8%
Weekend	Cent.	92.7%	89.5%	86.1%
	S-OFF1	90.2%	88.1%	78.6%
	S-OFF2	89.3%	86.6%	78.5%

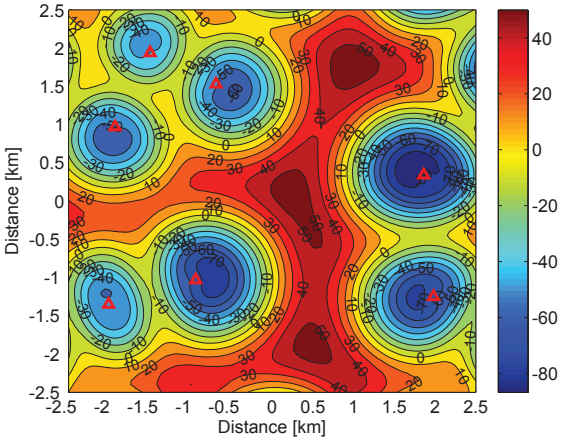
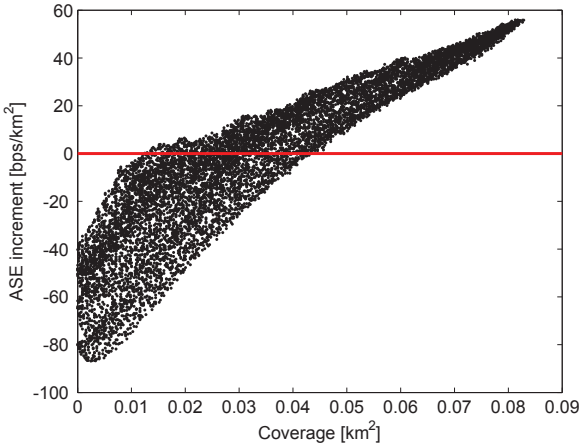
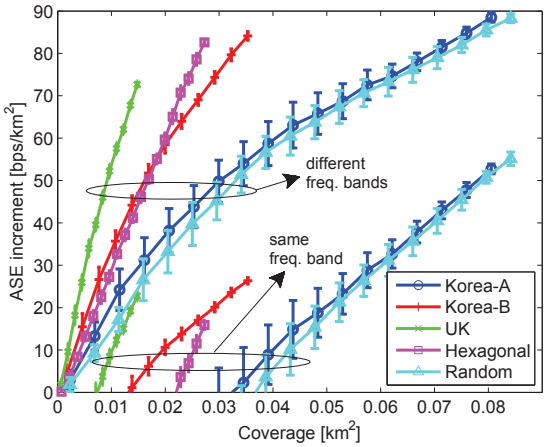


Figure 1: Contour plot of ASE increment (Korea-A).



(a) Scatter plot between the coverage and ASE increment (Korea-A)



(b) Quantile plot between the coverage and ASE increment

Figure 2: Several interesting observations from various topologies including the real layout of macro BSs.

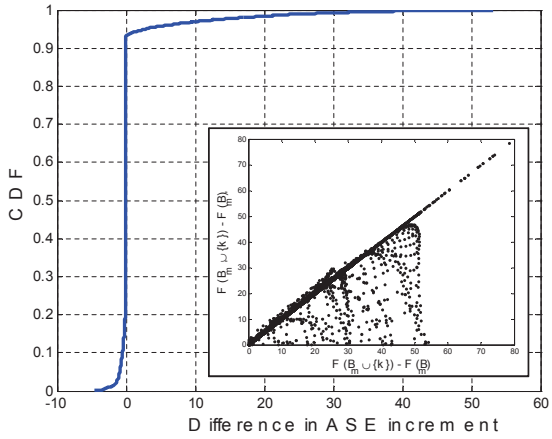


Figure 3: Submodularity test under the inhomogeneous environment.

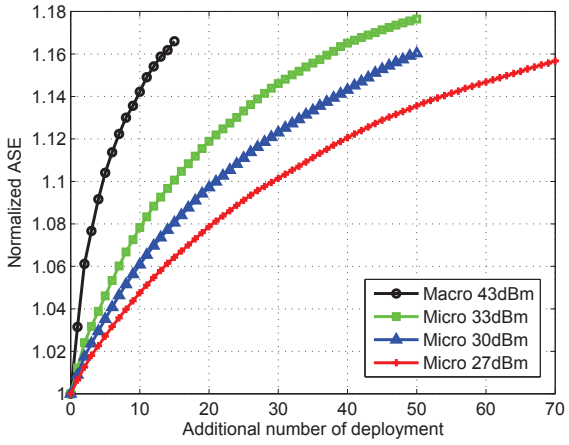


Figure 4: Normalized ASE increment according to the deployment of different types of BSs.

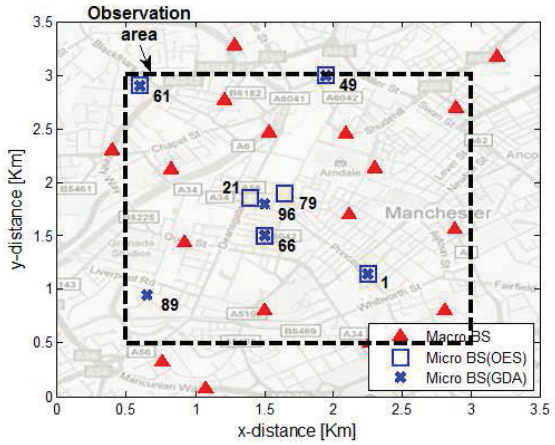


Figure 5: A snapshot after 6 micro BSs are additionally deployed by the OES and GDA on top of the real topology of macro BSs in Manchester, UK.

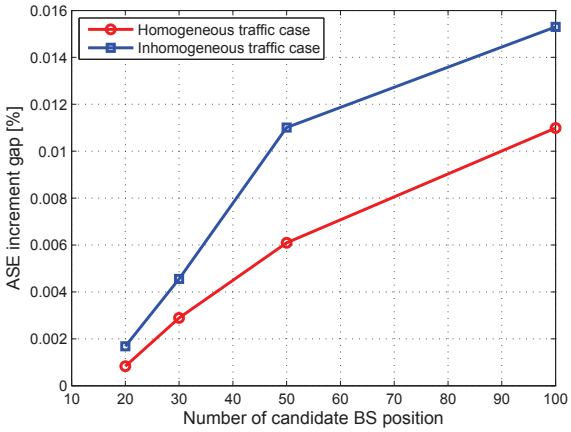


Figure 6: ASE increment gap between the OES and GDA under the homogeneous and inhomogeneous traffic. The performance is averaged over 20 random $\mathcal{K}$ generations.

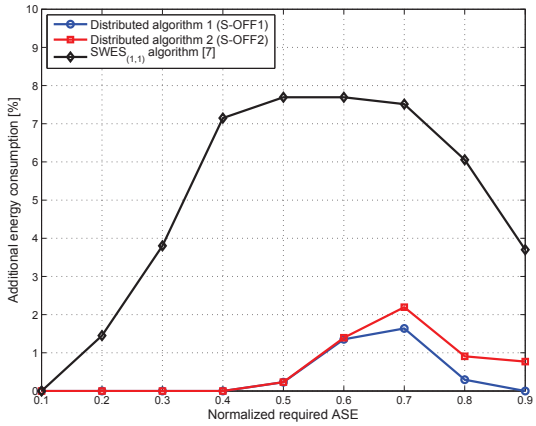


Figure 7: Percentage of additional energy consumption compared to that of the offline centralized algorithm.

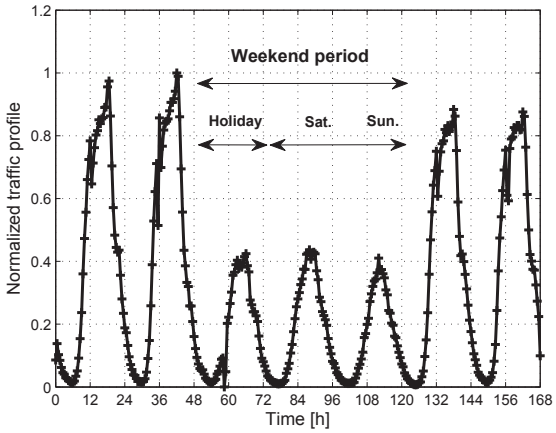


Figure 8: Normalized traffic profile during one week from a real cellular data trace [5].