Energy Cost Reduction in Cellular Networks through Dynamic Base Station Activation

Ali Abbasi and Majid Ghaderi
Department of Computer Science, University of Calgary
{aabbasi, mghaderi}@ucalgary.ca

***Abstract*—In this paper, we investigate dynamic base station activation with the aim of reducing energy consumption in cellular networks. Using the two-timescale *Lyapunov optimization* approach, we develop an online control algorithm to choose active set of base stations so as to satisfy users' demands while incurring minimum energy consumption. The algorithm selects the minimum cardinality subset of base stations that ensures stabilization of user queues. Our algorithm achieves stabilization without relying on instantaneous feedback about the network conditions, instead it only requires information about the average load and demand over a coarse time scale. The formulated problem which consists of joint base station activation and user association is generally intractable. However, we show that it features *submodularity*, and consequently present a near-optimal solution for certain instances of the problem. We further develop a greedy algorithm to solve general cases of the problem. We supplement our theoretical analysis with numerical results to demonstrate the behavior of our algorithm in terms of energy and delay in some example network scenarios.**

I. INTRODUCTION

There has been a massive growth in cellular data traffic over the past few years. This trend is expected to continue and has already resulted in a global shortage of wireless bandwidth [1]. To accommodate increasing volumes of traffic, mobile operators are moving towards denser deployment of base stations (BSs) in order to provide more capacity by increasing the spatial reuse of radio frequencies [2]. In a dense deployment, each base station covers a small geographical area and serves a small number of users, which allows it to provide them with higher rate. While this is an effective approach to provide better service to users, deploying a large number of base stations could result in a significant increase in network energy consumption. This has become a great concern for the cellular operators due to higher operational expenditure, *e.g.*, electricity bill, as well as higher carbon footprint. Therefore, energy efficiency has emerged as a critical performance metric for cellular networks [3].

As reported in [4], among various elements of a cellular network, base stations account for 60-80% of the total network energy consumption. A notable observation is that while cellular traffic exhibits periodic behavior, the energy consumption approximately stays the same [5]. This can be attributed to the fact that cellular operators often deploy as many base stations as necessary to satisfy the *peak* traffic demand, while keeping them active (*i.e.*, in the On state) all the time. In addition, due to various sources of energy consumption in BS equipment (*e.g.*, cooling system and processing unit), transmission power control mechanisms alone cannot compensate for BSs being always active. Particularly, with current base stations, about 50-90% of peak energy (energy consumed during the peak traffic) is consumed even in idle or low traffic state [3]. Dynamic base station activation has been proposed and considered as a viable solution to address this problem [6], [7].

The idea is to completely power off underutilized base stations when their traffic load could be handled by nearby base stations, and in turn, power on some inactive base stations when the load in their coverage area exceeds the capacity of the current active BSs, so as to satisfy the demand. It has been observed that by dynamically activating base stations in a network, significant energy savings can be achieved [5]. In this paper, we investigate this idea with the aim of minimizing the *long-term energy cost* of operating a cellular network. This is a challenging problem as solving it requires knowledge of future network conditions, *e.g.*, traffic load and power price. Since this information is not available a priori, we seek *online* control mechanisms which do not rely on such information, as they utilize the knowledge of current network conditions to minimize the long-term energy consumption.

To this end, we model the problem following the framework of stochastic optimization. An important feature of our approach is that it relies only on the information that is readily available in current cellular networks, *e.g.*, information about user queue backlogs. Utilizing the recent results from the two-timescale Lyapunov optimization technique [8], we formulate the problem and specify the control decisions that the system implements in order to minimize the long-term energy consumption, while stabilizing user queues. Our algorithm only requires the knowledge of the average data rate supported by each base station and the average traffic arrival rate for each user. As a result, our algorithm adds minimal overhead to the backhaul links connecting base stations to the core network elements, where the control algorithm resides. At a longer time scale (*e.g.*, minutes), the controller decides about the set of active base stations, while decisions that require a shorter time scale (*e.g.*, milli-seconds) such as user association and transmission power control are delegated to base stations themselves.

This work is supported in part by the Smart Applications on Virtual Infrastructure (SAVI) project funded under the National Sciences and Engineering Research Council of Canada (NSERC) Strategic Networks grant number NETGP394424-10.

978-1-4799-4657-0/14/$31.00

There are some recent works on the subject. In [5], a location-dependent traffic profiling study is conducted on real 3G network traces showing that 23-53% energy saving is possible via dynamic base station activation. Operators cooperation is investigated in [9], where the optimal switch-off frequencies of base stations are computed in order to achieve balanced energy savings and roaming costs. Assuming a sinusoidal traffic profile, and following a threshold-based activation rule by each base station, an analysis of achievable energy savings is provided in [7]. The closest works to our work are presented in [10] and [6]. In [10], centralized and heuristic methods are presented for finding and deactivating the base station with the lowest load. The joint problem of base station activation and user association is studied in [6], where the objective is to minimize a joint energy and delay cost function. Our work differs from the above mentioned works in several aspects: 1) we consider the long-term energy cost of the system as opposed to short-term energy cost, 3) we systematically incorporate queue backlogs into our formulation which allows to control the relative importance of delay versus energy cost 3) our algorithm operates over a long timescale that causes minimal backhaul overhead, and 4) our algorithm only requires information about the average traffic and demand rate.

The two-timescale Lyapunov optimization for the purpose of power reduction in data centers has been employed in [8]. However, unlike [8], the problem in our work has a combinatorial nature (due to On/Off behavior of base stations) which makes it different and more challenging to solve. The optimal sleep-wake scheduling for energy harvesting mobile devices using Lyapunov optimization has recently been considered in [11]. The work investigates On/Of scheduling for only one device, while we consider such scheduling over a set of base stations.

Our contributions in this work can be summarized as follows:

- We formulate the base station activation problem as a stochastic optimization problem aiming to minimize the long-term energy cost of the cellular network.
- We derive a control decision problem based on the two-timescale Lyapunov optimization technique by employing a suitable Lyapunov function and deriving an upper bound on its T-slot drift function.
- We show that the main decision problem is submodular and present efficient approximation algorithms to solve it. For certain cases, we present a $1/3$-optimal approximation algorithm.

The rest of the paper is organized as follows. In Section II, the system model and formulation of the energy cost minimization problem are presented. In Section III, the Lyapunov framework to solve the problem is introduced and the corresponding control problem is derived. The approximation algorithms are presented in Section IV. Sample numerical results are presented in Section V. Section VI concludes the paper.

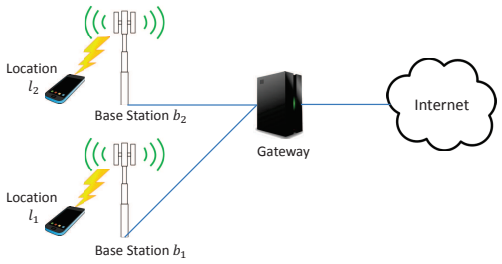


Fig. 1. A set of base stations is connected to a gateway.

II. System Model and Problem Formulation

A. Network Model

We consider a network similar to the one depicted in Fig. 1. The network consists of the gateway (system controller), the set $\mathcal{B} = \{b_1, \ldots, b_n\}$ of base stations that jointly provide coverage for the set $\mathcal{L} = \{l_1, \ldots, l_m\}$ of locations [1]. In our model, instead of dealing with individual users, we consider *locations*, where multiple users can be present in a single location. The system graph is defined as the bipartite graph $G_s = (\mathcal{B} \cup \mathcal{L}, \mathcal{E})$ in which an arc $e_{ij} \in \mathcal{E}$ connects BS b_i to location l_j if l_j is under the coverage of b_i (notion of coverage will be clarified later). Let $\mathcal{L}_i$ denote the subset of locations that are under the coverage of BS b_i and $\mathcal{B}_j$ denote the set of all base stations that cover location l_j.

BS activation decisions are made by the controller which operates on a discrete-time basis. To avoid unnecessary overhead due to transient network states, BS activation is performed at a timescale that is different from other network operations such as user association and scheduling. Therefore, time is divided into frames of size T timeslots. Activate and deactivate decisions are made at the beginning of each time frame. Binary vector $\boldsymbol{Y}(t) = [y_i(t)]_{|\mathcal{B}|}$ which is defined as follows

$$y_i(t) = \begin{cases} 1, & \text{If } b_i \text{ is active at } t, \\ 0, & \text{otherwise.} \end{cases}$$

denotes the set of active base stations at timeslot t.

B. Resource Allocation

Similar to the LTE systems [12], we consider an OFDMA-based radio access interface. In OFDMA systems, available frequency bandwidth is partitioned into orthogonal subcarriers. The transmit power P is divided between these subcarriers.

Assume that base station b_i communicates with a user at location l_j. Let p_{ij} denote the power allocated to l_j from b_i. Let g_{ij} denote the power gain between b_i and l_j received signal power of the user is given by $g_{ij} \cdot p_{ij}$. A location is considered covered by a base station if the received power of the pilot signal at that location is higher than a prespecified threshold. The achievable rate of a user has direct relation to the received

[1] Such a network model is consistent with 4G cellular networks based on LTE technology [12].

Signal-to-Noise-and-Interference Ratio (SINR). SINR of the received signal at l_j when served from BS b_i is given by

$$\text{SINR}_{ij} = \frac{g_{ij}p_{ij}}{n_j + I_j}, \tag{1}$$

where n_j and $I_j = \sum_{b_{i'} \in \mathcal{B}_j \setminus b_i} g_{i'j}p_{i'j}$ are background noise power and interference power at location l_j, respectively.

The achievable rate of the user is obtained from a rate function $R(\cdot)$, which is generally increasing and concave w.r.t SINR_{ij}. A common choice is the Shannon capacity formula. Similar to [13], to make the formulation tractable, we use an upper bound I on the interference power instead of using an exact expression for I_j. Following [14], I denotes the maximum multi-cell interference level. While this approximation results in a conservative estimation of the achieved rate, it does not need substantial amount of signaling to compute the actual value of interference at each location. We further borrow the following simplifying assumptions from the literature [15], [16]:

- In OFDMA systems *e.g.*, LTE, each subcarrier is shared among multiple users using TDM *i.e.*, resources are shared in both frequency and time. We assume that each subcarrier can be fractionally shared among users [15]. This assumption is particularly true in our system as we consider the long-term averages during frames.
- We assume the total transmission power P is divided equally among all subcarriers. Therefore, if the bandwidth is partitioned into R subcarriers, the allocated power to each one is $p = P/R$. While it is possible to consider optimal power allocation across subchannels, it is a problem that is orthogonal to the problem considered in this paper. Moreover, as shown in [15], this scheme is nearly-optimal and is widely used in practice.
- As provisioned in LTE networks [16], we assume that neighboring cells are able to coordinate allocation of resources to users in overlapping regions such that orthogonal resources are allocated from neighboring BSs to locations in overlapping regions.

In the rest of the paper, we use the term *resources* to refer to subcarriers. Let $0 \leq \gamma_{ij} \leq 1$ denote the fraction of resources allocated to location l_j from base station b_i. Following the third assumption, the total received rate at location l_j is given by:

$$r_j = \sum_{b_i \in \mathcal{B}_j} \gamma_{ij} \cdot R_{ij}, \tag{2}$$

where R_{ij} is the rate received from base station b_i if all of its resources were to be allocated to location l_j, *i.e.*,

$$R_{ij} = R \cdot \log(1 + \beta \frac{pg_{ij}}{n_j + I}). \tag{3}$$

In (3), β is the SINR gap due to limited modulation levels available in practice. Let $\mathcal{R}_{\boldsymbol{Y}}$ denote *rate region* of $\boldsymbol{Y}$ *i.e.*, the set of all rate vectors $\boldsymbol{r} = [r_j]_{|\mathcal{L}|}$ achievable at all locations $\mathcal{L}$ by the active base stations $\boldsymbol{Y}$. We then have

$$\mathcal{R}_{\boldsymbol{Y}} = \Big\{ \boldsymbol{r} = [r_j] : \sum_{b_i : y_i = 1} \gamma_{ij} R_{ij} = r_j, \sum_{l_j \in \mathcal{L}_i : y_i = 1} \gamma_{ij} \leq 1 \Big\}. \tag{4}$$

C. Energy Cost Model

If an active BS b_i at timeslot t consumes total power P_{b_i}, then the energy cost incurred by b_i is given by

$$C_i(t) = C_P(t) \cdot P_{b_i}(t), \tag{5}$$

where $C_p(t)$ is the energy price at t. $C_p(t)$ changes according to an exogenous random process which is assumed to have a stationary distribution. The total power consumption $P_{b_i}(t)$ consists of two parts [5] as follows:

$$P_{b_i}(t) = P_{tx}(t) + P_{misc}(t), \tag{6}$$

where $P_{tx}(t)$ is the transmission power used to communicate with users at t. The term $P_{misc}(t)$ accounts for the base power spent in cooling and power supply at t. Clearly, P_{tx} depends on the carried load traffic and can be approximated as follows [5],

$$P_{tx}(t) = P_\alpha \cdot \mu(t) + P_\beta, \tag{7}$$

where $\mu(t)$ is the traffic load factor of the base station at t. The slope and offset power coefficients P_α and P_β are constants that vary for different types of base stations from different vendors. Overall, there is a base cost for activating a base station due to residual factors P_{misc} and P_β and a traffic dependent part due to P_α. As reported in [17], the base activation cost can take up to 50% of the total base station power consumption.

D. Problem

The traffic intended for users is first received at the gateway and stored in user queues. The gateway keeps queue $Q_j(t)$ for each location l_j[2]. We denote the amount of workload arrived at timeslot t as $\boldsymbol{A}(t) = [A_1(t), \ldots, A_m(t)]$. We assume the arrival at location l_j follows an i.i.d. distribution throughout the whole frame while the average rate $\bar{A}_j$ is known to the gateway (the gateway can estimate this over each frame). In addition, we assume that there exist bounds A_{min} and A_{max} such that $A_{min} \leq A_j(t) \leq A_{max}$ for all $l_j \in \mathcal{L}$.

The data stored in queue Q_j will be disseminated among all the base stations that provide service to location l_j. Let $\mu_{ji}(t)$ denote the amount of l_j's traffic routed to BS b_i at the beginning of timeslot t. We assume that the system tries to match the amount of data transferred to BS b_i to the service rate provided to l_j by b_i at timeslot t *i.e.*,

$$\mu_{ji}(t) \leq \gamma_{ij}(t) \cdot R_{ij}(t) \qquad \forall b_i \in \mathcal{B}, l_j \in \mathcal{L}, \tag{8}$$

while,

$$\sum_{l_j \in \mathcal{B}_i} \gamma_{ij}(t) \leq 1 \qquad \forall b_i \in \mathcal{B}. \tag{9}$$

[2]In our method, we only need information regarding the backlog state of each location. In practice, queues might actually be resided in BSs while information regarding their sizes is fed back to the gateway.

Although in practice, base stations keep buffers for the data that are not transmitted yet, this assumption simplifies the queuing model considered in this paper. In addition it does not affect the resulting solution as we make decisions based on the whole backlog for each location. Obviously, if b_i does not cover l_j then $\mu_{ji}(t) = 0$. We assume that the rate provided for each location is subject to maximum bound μ_{max} such that the inequality $0 \leq \sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(t) \leq \mu_{max}$ holds for all locations $l_j \in \mathcal{L}$. Queues evolve in consecutive timeslots according to the following queuing dynamic,

$$Q_j(t+1) = \max[Q_j(t) - \sum_{b_i \in \mathcal{B}} y_i(f_t)\mu_{ji}(t), 0] + A_j(t)\,. \quad (10)$$

We say that the system is stable if the following condition holds on queue backlogs,

$$\overline{Q} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{j=1}^{m} \mathbb{E}\{Q_j(\tau)\} < \infty\,. \quad (11)$$

The energy cost of the system at timeslot t is the sum of the energy cost of all base stations, *i.e.*,

$$Cost(t) = \sum_{b_i \in \mathcal{B}} C_i(t)\,. \quad (12)$$

The problem we consider in this paper is to minimize the long-term cost of the system defined as follows,

$$\textbf{P1: Minimize } \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} Cost(\tau) \quad \text{subject to: (8), (9), (11)}\,. \quad (13)$$

In the next section, we derive a control algorithm to solve (13).

III. ONLINE ALGORITHM DESIGN

To develop the online control algorithm, we first define the Lyapunov function $L(t)$ as a scalar measure of queue backlog in the system as follows,

$$L(t) \triangleq \sum_{j=1}^{m} \frac{1}{2}[Q_j(t)]^2\,. \quad (14)$$

It is desirable for our algorithm to push the system towards a lower backlog state. Therefore, to observe the expected change in the Lyapunov function over T timeslots, we define the T-slot Lyapunov drift as follows,

$$\Delta_T(t) \triangleq \mathbb{E}\{L(t+T) - L(t)|\boldsymbol{Q}(t)\}\,. \quad (15)$$

In addition, we would like to minimize the long-term energy cost of the system as defined in (13). Hence, following the *drift-plus-penalty* approach [18], we add the expected energy cost of the system to (15), which results in the following drift-plus-penalty expression,

$$\Delta_T(t) + V \cdot \mathbb{E}\{\sum_{\tau=t}^{t+T-1} Cost(\tau)\}, \quad (16)$$

where the parameter V is chosen so as to control the trade-off between energy cost and congestion (reflected in queue backlogs) in the network. The next derivation step in Lyapunov optimization is to find an upper bound on this expression. We show that the following theorem holds.

Theorem 1. *Let $V > 0$ and $t = kT$ for some $k \in \mathbb{Z}_+$. For any set of possible activation decisions $\boldsymbol{Y}(t)$ and user associations $\boldsymbol{\mu}(t)$, we have,*

$$\begin{aligned}\Delta_T(t) &\leq BT \\ &- \mathbb{E}\{\sum_{\tau=t}^{t+T-1} \sum_{l_j \in \mathcal{L}} Q_j(\tau)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big]|\boldsymbol{Q}(t)\} \\ &+ V\mathbb{E}\{\sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\},\end{aligned}$$

where, $B = \frac{1}{2}m(A_{max}^2 + \mu_{max}^2)$.

Proof: Assume $\tau \in [t, t+T-1]$. Squaring the queuing dynamics (10), the following inequality is obtained,

$$\begin{aligned}Q_j(\tau+1)^2 \leq &Q_j(\tau)^2 + [\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau)]^2 + A_j(\tau)^2 \\ &-2Q_j(\tau)\big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau)\big] + 2A_j(\tau)Q_j(\tau)\,.\end{aligned} \quad (17)$$

Summing (17) over all locations $l_j \in \{l_1, \ldots, l_m\}$ and using inequalities $\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) \leq \mu_{max}$ and $A_j(\tau) \leq A_{max}$, we have,

$$\begin{aligned}\frac{1}{2}\sum_{l_j \in \mathcal{L}} \big[Q_j(\tau+1)^2 - Q_j(\tau)^2\big] &\leq \frac{1}{2}m(A_{max}^2 + \mu_{max}^2) \\ &- \sum_{l_j \in \mathcal{L}} Q_j(\tau)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big]\,.\end{aligned}$$

By taking the expectation of both sides w.r.t to arrival traffic from BSs to locations conditioned on $\boldsymbol{Q}(t)$, it is obtained that,

$$\begin{aligned}\Delta_1(\tau) &\leq B - \\ &\mathbb{E}\{\sum_{l_j \in \mathcal{L}} Q_j(\tau)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big]|\boldsymbol{Q}(t)\}\,.\end{aligned} \quad (18)$$

Summing (18) over $\tau = [t, \ldots, t+T-1]$ and adding the cost term $V\mathbb{E}\{\sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\}$, yields the following:

$$\begin{aligned}\Delta_T(t) &\leq BT \\ &- \mathbb{E}\{\sum_{\tau=t}^{t+T-1} \sum_{l_j \in \mathcal{L}} Q_j(\tau)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - R_j(\tau)\Big]|\boldsymbol{Q}(t)\} \\ &+ V\mathbb{E}\{\sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\}\,.\end{aligned} \quad (19)$$

■

The rule of Lyapunov optimization is to choose the control action so as to minimize the right-hand side of (19). To do so, we need information on the queue backlogs $Q_j(\tau)$ in timeslots $\tau = t, t+1 \ldots, t+T-1$, which is not available at the beginning of the frame which is t. Therefore, we approximate

the queue backlog at each timeslot τ, *i.e.*, $Q_j(\tau)$, by the queue backlog at the beginning of the frame, *i.e.*, $Q_j(t)$. However, doing so loosens the upper bound obtained in (19) as explained in the following. From queuing dynamics (10), the following inequality holds for every timeslot $\tau \in [t, t+T-1]$,

$$Q_j(t) - (\tau - t)\mu_{max} \leq Q_j(\tau) \leq Q_j(t) + (\tau - t)A_{max}\,.$$

Therefore, from (19), we obtain that,

$$\begin{aligned}\Delta_T(t) \leq BT - \mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{l_j \in \mathcal{L}} \big[Q_j(t) - (\tau - t)\mu_{max}\big] \\ \times \Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big] | \boldsymbol{Q}(t)\} \\ + V\mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\},\end{aligned}$$

which, leads to the following expression,

$$\begin{aligned}\Delta_T(t) \leq BT - \\ \mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{l_j \in \mathcal{L}} Q_j(t)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big] | \boldsymbol{Q}(t)\} \\ + \frac{T(T-1)}{2} m\mu_{max}[\mu_{max} - A_{min}] \\ + V\mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\}\,.\end{aligned}$$

Define $B_1 = B + \frac{T-1}{2} m\mu_{max}[\mu_{max} - A_{min}]$. It follows that,

$$\begin{aligned}\Delta_T(t) \leq B_1 T - \\ \mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{l_j \in \mathcal{L}} Q_j(t)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big] | \boldsymbol{Q}(t)\} \\ + V\mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\}\,.\end{aligned} \tag{20}$$

In the next section, we show how this expression can be used to design our control algorithm.

IV. Solution

Our goal is to minimize the R.H.S. of (20) or equivalently maximize the following term,

$$\begin{aligned}\mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{l_j \in \mathcal{L}} Q_j(t)\Big[\sum_{b_i \in \mathcal{B}} y_i(t)\mu_{ji}(\tau) - A_j(\tau)\Big] | \boldsymbol{Q}(t)\} \\ - V\mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)C_i(\tau)\}\,.\end{aligned} \tag{21}$$

Knowing the average arrival rate for each location j, the term (21) is reduced to the following,

$$\mathbb{E}\{ \sum_{\tau=t}^{t+T-1} \sum_{b_i \in \mathcal{B}} y_i(t)\big[\sum_{l_j \in \mathcal{L}} Q_j(t)\mu_{ji}(\tau) - VC_i(\tau)\big]\}\,. \tag{22}$$

Using (22), the optimization problem now can be stated as the joint optimization of user association (determining $\mu'_{ij}s$) at each timeslot and base station activation (finding optimal activation vector $\boldsymbol{Y}$) at the beginning of each frame.

Assume that initially all the queues are empty. For a given $V > 0$, It has been shown [19] that any method that maximizes (22) satisfies the following

$$\limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} Cost(\tau) \leq Cost^* + \frac{D}{V} \tag{23}$$

where $Cost^*$ is the minimal achievable long-term cost that stabilizes queues under any activation policy and D is a constant. This indicates that by increasing V the distance between the achieved cost and the optimal one can be made arbitrarily small. Note that larger V means larger queue sizes as we usually have performance-backlog trade-off [19].

To solve the problem (22), we need to make a few simplifying assumptions. Recall that reducing the number of active base stations will reduce energy cost. Therefore, we consider the worst case energy cost when activating a BS. Particularly, we assume that when a BS is activated, it transmits at maximum transmission power. Fully utilizing the active BSs seems to be the best policy as base power consumed for cooling and idle mode signaling is comparable with its transmission power consumption. Thus, in (22), $C_i(\tau)$ is replaced with constant C_i such that $C_i(\tau) \leq C_i$. Note that by reducing the number of active BSs, we lean more towards energy cost reduction and may increase queue sizes in the network. On the other hand, the amount of backlog in the system can be controlled by choosing the right value of control parameter V.

In addition, recall that to compute the optimal user associations, only the queue backlog information at the beginning of a frame is used in (22). This fact along with the assumption activation decisions do not change during the frame indicates that, there is an optimal solution to (22) in which associations between base stations b_i and locations l_j, *i.e.*, $\bar{\mu}_{ji}$, are constant.

Taking the above assumptions into consideration, problem (22) is transformed to the following problem,

$$\begin{aligned}\textbf{P2: Maximize } N(\boldsymbol{Y}) = \sum_{b_i \in \mathcal{B}} y_i\big[\sum_{l_j \in \mathcal{L}} Q_j\bar{\mu}_{ji} - VC_i\big] \\ \text{Subject to: } \bar{\mu}_{ji} \leq \gamma_{ij}R_{ij}, \quad \forall l_j \in \mathcal{L}, b_i \in \mathcal{B} \\ \sum_{l_j \in \mathcal{L}_i} \gamma_{ij} \leq 1, \quad \forall b_i \in \mathcal{B} \\ \sum_{b_i \in \mathcal{B}_j} \gamma_{ij}R_{ij} \leq Q_j,\end{aligned} \tag{24}$$

$N(\boldsymbol{Y})$ is called the *net utility* of the system. Q_j denotes the queue backlog of location l_j at the beginning of the current time frame. The rate that transfered to BS b_i intended for location l_j *i.e.*, $\bar{\mu}_{ji}$ cannot be larger than the rate supported by the BS *i.e.*, $\gamma_{ij}R_{ij}$. R_{ij} is computed based on long-term average power gain between b_i and l_j. Moreover, rate allocated to l_j cannot exceed its demand *i.e.*, Q_j. **P2** belongs to the class of maximum facility location [24] problems which are generally NP-hard.

In the following, based on the concept of generalized network flows [20], we demonstrate that Problem **P2** is a nonmonotone submodular maximization problem [21]. This property allows us to employ approximation algorithms proposed to solve these types of problems [21], [22]. The demonstration is through the decomposition of the objective into two joint goals. The first part is to choose active base stations and associate users to them so as to maximize the following term,

$$\sum_{b_i \in \mathcal{B}} y_i \big[\sum_{l_j \in \mathcal{L}} Q_j \bar{\mu}_{ji} \big], \tag{25}$$

which is the sum of flows from BSs to users. The second part is to choose active base stations so as to minimize the following term,

$$\sum_{b_i \in \mathcal{B}} y_i V C_i, \tag{26}$$

which is the energy cost of active base stations. In the next sections, the concepts of submodular functions and generalized flows are introduced in order to solve this problem.

A. Submodular Functions

Submodular functions are discrete counterparts of convex/concave functions[3]. A set function $f(.)$ defined over the ground set $\mathcal{V}$ is submodular if it satisfies the following property for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \backslash \{v\}$,

$$f(\mathcal{A}+v) - f(\mathcal{A}) \geq f(\mathcal{B}+v) - f(\mathcal{B}). \tag{27}$$

This property is called the *diminishing return* property as it states that adding an element to a smaller context would make more difference in the function value than adding it to a larger context. If negation of (27) holds for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \backslash \{v\}$, then the function is called supermodular. A set function f is monotone if $f(\mathcal{A}) \leq f(\mathcal{B})$ for all $\mathcal{A} \subseteq \mathcal{B}$. If f and g are submodular functions then $\alpha f + \beta g$ is submodular for any $\alpha, \beta \geq 0$. A linear (modular) function f is defined as $f(\mathcal{A}) = \sum_{i \in \mathcal{A}} w_i$ for some weight function $w : V \rightarrow \mathbb{R}$. Linear functions are both submodular and supermodular.

B. Generalized Maximum Flow

Generalized network flow problems are similar to the traditional network flow problems except that a link gain is defined on each link. Let $G = (\mathcal{V}, \mathcal{E})$ be a directed graph. Let $u_{vw} \geq 0$, $\gamma_{vw} > 0$, and f_{vw} denote capacity, gain, and flow on link (v, w). The received flow at w from v will be given by $\gamma_{vw} \cdot f_{vw}$. Each node v has excess d_v. If $d_v > 0$, v is a source. There also exists a sink node t. The maximum flow problem for G is an optimal solution to the following linear program:

$$\begin{aligned} \text{Maximize} \quad & \sum_{v} \gamma_{vt} f_{vt}, \\ \text{subject to:} \quad & \sum_{w} f_{vw} - \gamma_{wv} f_{wv} \leq d_v, \quad \forall v \in \mathcal{V} \\ & f_{vw} \leq u_{vw}, \quad \forall (v,w) \in \mathcal{E} \\ & f_{vw} \geq 0, \quad \forall (v,w) \in \mathcal{E}\,. \end{aligned} \tag{28}$$

The first constraint ensures that the flow out of a source does not exceed its capacity. Suppose there exists a set function that maps every subset of $\mathcal{V}$ to the maximum generalized flow obtained from only the sources included in that subset. It has been shown that the this function is submodular [20].

To show that the maximum base station flow problem (25) is submodular, we extend the system graph G_s defined in Section II as follows. A virtual sink t is added to the set of nodes. Also arcs (j, t) are added to connect every location l_j to t. The capacity of each arc (j, t) and its gain are both set to the queue backlog of location l_j, *i.e.*, Q_j. In addition, the excess from each base station is assumed to be 1 which represents the whole fraction of resources that can be allocated to users. Gain and capacity of each arc connecting BS b_i to location l_j are set to R_{ij}. It follows that the maximum generalized flow on the constructed graph is equal to the optimal flow obtained from (25). Based on the results in [20], this demonstrates that maximum BS flow problem is submodular.

Moreover, the energy cost component of the Problem **P2** (24) is linear and hence supermodular. Since $\sum_{b_i \in \mathcal{B}} V C_i$ is supermodular, $-\sum_{b_i \in \mathcal{B}} V C_i$ is submodular, which means that (24) is submodular due to the closure of submodular functions under the addition operation.

C. Approximate Solutions

We first assume that the net utility is nonnegative for all possible subsets of $\mathcal{B}$ that are active, though it might be nonmonotone. This happens when the costs of activating BSs are small compared to the flow rates that they provide to users. Then, **P2** (24) problem is an instance of maximizing a nonnegative nonmonotone submodular function. Hence, to solve it, we can employ general combinatorial solutions to this type of problems as presented, for example, in [21]. The algorithm presented here (see Algorithm 1) is based on the work by Buchbinder *et al.* [22] that has a very simple structure and guarantees finding a $\frac{1}{3}$-optimal solution.

The algorithm starts with sets $\mathcal{X}$ and $\mathcal{Y}$ initialized to $\emptyset$ and $\mathcal{B}$, respectively. Every BS $b_i \in \mathcal{B}$ is considered and both sets agree on its inclusion in the final solution. If the utility (added flow minus the BS cost) of adding the BS to set $\mathcal{X}$ is greater than the lost utility of removing it from $\mathcal{Y}$, it will be added to $\mathcal{X}$. Otherwise, it will be removed from $\mathcal{Y}$. At the end, both sets are identical and give the set of the base stations to be activated. The algorithm runs $O(n)$ number of times, scanning all the BSs. In each iteration, to obtain the added utility, a linear program similar to (28) is solved. The running time of the algorithm is the product of $O(n)$ and the running time of solving the linear program which is dependent on the type

[3]They share properties of both types of functions. Similar to convex functions, their minimum value can be obtained in polynomial time while finding the maximum is NP-hard. Also similar to concave functions, they feature diminishing return property.

of linear optimization method employed. For example, linear optimization can be carried out in $O(\frac{n^3}{\ln n}S)$ [23] where S is the bit length of data and n is dimension of the optimization vector.

Algorithm 1: Set-Matching Activation Algorithm

Input: $\mathcal{B}$
Output: $\boldsymbol{Y}$
begin
  $\mathcal{X} \leftarrow \emptyset$;
  $\mathcal{Y} \leftarrow \mathcal{B}$;
  foreach $b_i \in \mathcal{B}$ **do**
    $r_x \leftarrow N(\mathcal{X} \cup b_i) - N(\mathcal{X})$;
    $r_y \leftarrow N(\mathcal{Y} \backslash b_i) - N(\mathcal{Y})$;
    if $r_x \geq r_y$ **then**
      $\mathcal{X} \leftarrow \mathcal{X} \cup b_i$;
    else
      $\mathcal{Y} \leftarrow \mathcal{Y} \backslash b_i$;
  $y_i \leftarrow 1, \quad \forall b_i \in \mathcal{X}$;

For instances of **P2** that nonnegativity property cannot be assumed *e.g.*, in sparse networks or when the cost of activating a BS is very high, the above algorithm cannot provide any optimzality guarantee. In fact when this is the case, non-monotone submodular maximization is inapproximable [24]. To deal with these situations, a greedy algorithm is presented in Algorithm 2. The intuition behind the algorithm is to make the best possible decision in each iteration.

The algorithm starts with an empty activation set $\boldsymbol{Y} = \boldsymbol{0}$. At each step, the base station that offers the highest utility is chosen and added to the set of active BSs. This process continues until no BS can be found that provides positive utility for the current set of active BSs. As one base station is activated in each iteration, the loop is executed at most n times. In addition, selection of the optimal base station in each iteration needs solving at most n linear programs, thus greedy activation needs to solve $O(n^2)$ linear programs to obtain the result. Combining this with the complexity of solving each linear program gives the the running time of Algorithm 2.

Algorithm 2: Greedy Activation Algorithm

Input: $\mathcal{B}$
Output: $\boldsymbol{Y}$
begin
  $\boldsymbol{Y} \leftarrow \boldsymbol{0}$;
  $continue \leftarrow true$;
  while *continue* **do**
    $b_{i^*} \leftarrow \arg\max_{b_i \in \mathcal{B}, y_i \neq 1} N(\boldsymbol{Y} \cup y_i) - N(\boldsymbol{Y})$;
    $u \leftarrow N(\boldsymbol{Y} \cup y_{i^*}) - N(\boldsymbol{Y})$;
    if $u > 0$ **then**
      $y_{i^*} \leftarrow 1$;
    else
      $continue \leftarrow false$;

V. Numerical results

In this section, we conduct a numerical study to better understand the properties of the proposed solution.

A. Setup

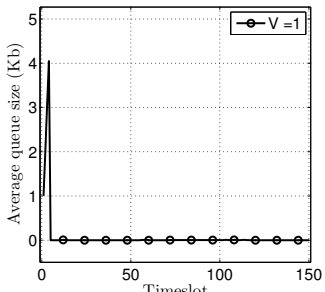
Transmission parameter values are adopted from [25] as the assumptions regarding channel gains are consistent with the standard 3GPP propagation models. The power gain between the sender and a receiver is $g = f(d)$ where d is the distance from the sender to the receiver in (km). $f(d) = 10^{h_0} d^{-\kappa}$ with path loss exponent $\kappa = 3.5$ and $h_0 = -14.4$. The background noise is $N_0 = -174$ dbm (Hz^{-1}). The bandwidth is 1 MHz and maximum power is set to $P = 10W$.

We consider a network of size $1200m \times 1200m$. Base stations are placed on a regular grid. The distance between each two neighboring BSs is $200m$. There are in total 25 base stations in the network. Activation costs of all BSs are the same. A base station is able to cover users which are up to $350m$ away from it. Two scenarios for the distribution of users in the network are considered: *uniform* and *non-uniform*. In the uniform case, location of a user is chosen uniformly at random in the network. To distribute users non-uniformly, nine crowded regions are considered in the network. Each crowded region is a circle of radius $160m$. Users are divided equally among theses crowded regions and distributed uniformly within each region.

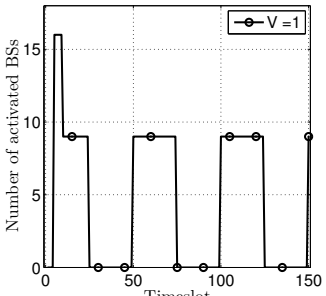
B. Energy vs delay tradeoff

Our goal is to study the behavior of the proposed algorithm in terms of energy cost and delay. These parameters can be approximated indirectly via the number of active BSs and average queue sizes respectively. Each frame is assumed to consist of 5 timeslots. Arrival to each location follows a Bernoulli process. Associated to each location l_j is a probability of acceptance p_j. This probability is determined randomly and independently for each location. With probability p_j, $40Kb$ is added to the queue of location l_j in each timeslot. To demonstrate how the algorithm responds to variation in the arrival traffic by the changing the set of active BSs, we divide frames into framesets. In even-numbered framesets arrived data goes to the gateway as normal, whereas there is no arrival in odd-numbered framesets. Each frameset consists of 5 frames. Cost of activating each BS is assumed to be 1100. For different values of control parameter V, average queue sizes and the number of active BSs are depicted in Fig. 2. As can be seen in the figure, initially the queues are empty and no BS is active. During the first frame, queues are filled up until the beginning of the next frame at which some base stations are activated. At this time, queues start to get empty. Overall, during even-numbered framesets, some of the BSs are active and queues have data. In odd-numbered, as there is no new arrival, gradually queues become empty and BSs are turned off. A notable observation is that by increasing V average queue sizes are increased. On the other hand, average number of active BSs is decreased. In addition, we see more variation both in queue sizes and the number of active BSs. In fact, with

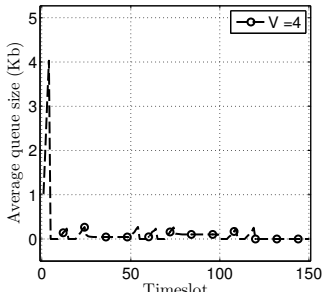
large V's, queues should become larger to justify activating BSs. However, when enough BSs are activated the system returns to the stable (smaller queues) state quite rapidly which in turn puts BSs into inactive state. In this set of results on average only 10 BSs are needed to satisfy the aforementioned arrival traffic demand. This shows great improvement in terms of energy consumption compared to activating all 25 BSs in the network.



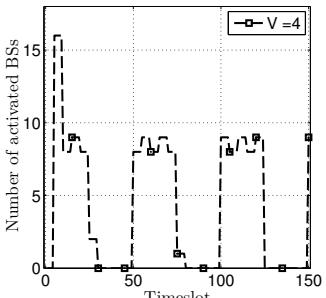
(a) Queue size variation, $V = 1$.



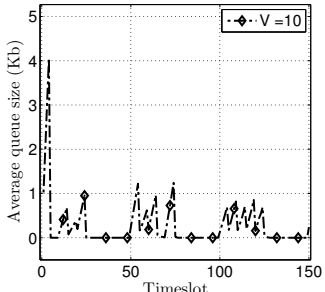
(b) Number of active BSs, $V = 1$.



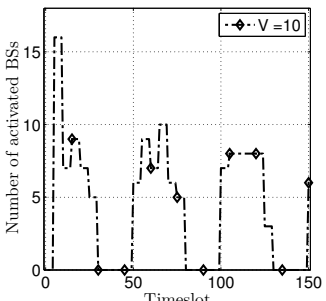
(c) Queue size variation, $V = 4$.



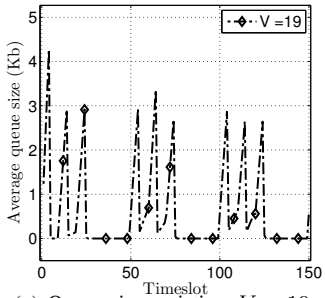
(d) Number of active BSs, $V = 4$.



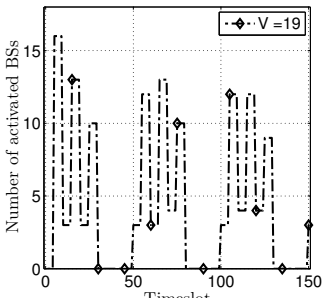
(e) Queue size variation, $V = 10$.



(f) Number of active BSs, $V = 10$.

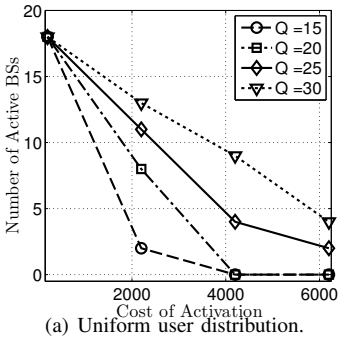


(g) Queue size variation, $V = 19$.

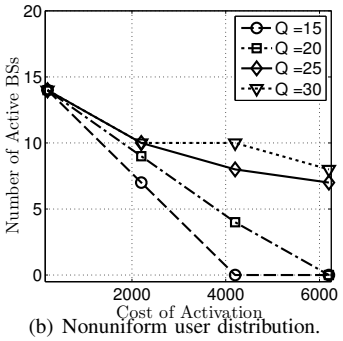


(h) Number of active BSs, $V = 19$.

Fig. 2. Variation of queues and active BSs versus V.



(a) Uniform user distribution.



(b) Nonuniform user distribution.

Fig. 3. Number active BSs varying activation cost.

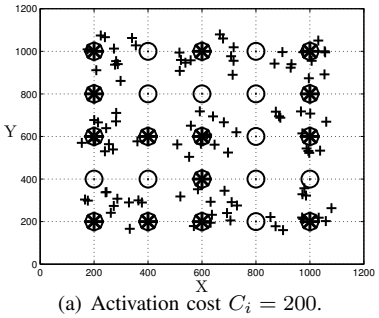
C. Effect of user distribution

Here, behavior of the algorithm is studied when distribution of users in the network is changed. To do so, for both cases of uniform and non-uniform user distribution, the number of activated base stations is determined varying both the activation cost and queue sizes. Fig. 3 shows the results. As can be seen in the figure, increasing the queues pushes more base stations to the active state while increasing the activation cost reduces the number of active base stations. In addition, uniform user distribution associates approximately the same number of locations to base stations which results in almost similar utilities for them. In comparison to the non-uniform user distribution, this allows activating more base stations when the cost is low. On the other hand, when the cost is high, activating almost every BS will result in negative utility which leads to lower number of active BSs compared to the non-uniform case.

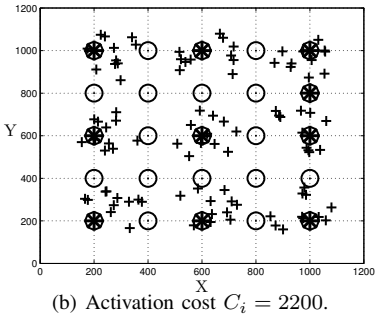
What is interesting, however, is the matching between base stations and crowded places. This is demonstrated in Fig. 4. This figure shows snapshots of the network and the corresponding active base stations while increasing the activation cost. The snapshots are taken from the above results where queue sizes are equal to $25Kb$. As apparent in the figure, base stations are activated in more populated areas which is expected according to the definition of utility.

VI. Conclusion and Future Work

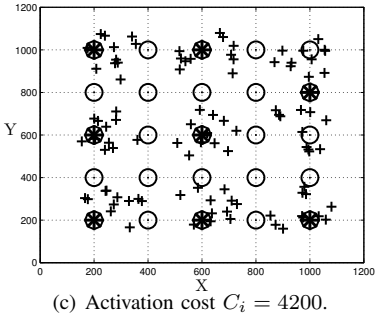
In this paper, we considered the dynamic base station activation problem with the objective of minimizing the long-term energy cost of the system. We proposed an online



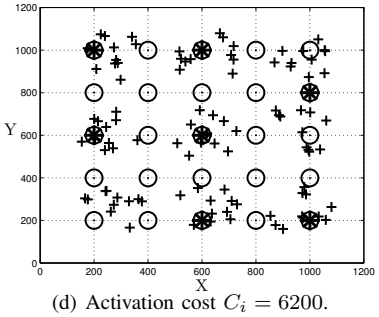
(a) Activation cost $C_i = 200$.



(b) Activation cost $C_i = 2200$.



(c) Activation cost $C_i = 4200$.



(d) Activation cost $C_i = 6200$.

Fig. 4. Snapshots of network and activated base stations for non-uniform user distribution ('$*$').

control algorithm by employing the two-timescale Lyapunov optimization technique. Our control algorithm does not rely on the full knowledge of system statistics except some that are readily available to the cellular network such as queue sizes. Our numerical results showed that the proposed control algorithm could deliver significant energy savings by dynamically activating the base stations.

REFERENCES

[1] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, 2013.

[2] J. Gong, S. Zhou, Z. Niu, and P. Yang, "Traffic-aware base station sleeping in dense cellular networks," in *Proc. IEEE IWQoS*, Beijing, China, Jun 2010, pp. 1–2.

[3] M. A. Marsan and M. Meo, "Green wireless networking: Three questions," in *Proc. IFIP Med-Hoc-Net*, Sicily, Italy, Jun 2011, pp. 41–44.

[4] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. IEEE ICC Workshops*, Dresden, Germany, Jun 2009, pp. 1–5.

[5] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3g cellular networks," in *Proc. ACM MobiCom*, Las Vegas, USA, Sep 2011, pp. 121–132.

[6] K. Son, H. Kim, Y. Yi, , and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, 2011.

[7] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proc. IEEE GlobeCom*, Miami, USA, Dec 2010, pp. 1–5.

[8] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, "Data centers power reduction: A two time scale approach for delay tolerant workloads," in *In Proc. IEEE INFOCOM*, Orlando, USA, Mar 2012, pp. 1431–1439.

[9] M. A. Marsan and M. Meo, "Energy efficient management of two cellular access networks," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 69–73, 2010.

[10] S. Zhouy, J. Gongy, Z. Yangy, Z. Niuy, and P. Yang, "Green mobile access network with dynamic base station energy saving," in *ACM MobiCom posters*, Beijing, China, Sep 2009.

[11] L. Huang, "Optimal Sleep-Wake Scheduling for Energy Harvesting Smart Mobile Devices," in *In Proc. WiOpt*, May 2013, pp. 484 – 491.

[12] "Long term evolution of the 3gpp radio technology," http://www.3gpp.org/LTE/.

[13] D. W. K. Ng and R. Schober, "Resource Allocation and Scheduling in Multi-Cell OFDMA Systems with Decode-and-Forward relaying," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2246–2258, 2011.

[14] FCC Spectrum Policy Task Force, "Report of the Spectrum Ffficiency Working," http://www.fcc.gov/sptf/reports.html, Nov. 2002.

[15] J. Huang, V. G. Subramanian, R. Agrawal, and R. A. Berry, "Downlink scheduling and resource allocation for ofdm systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 288–296, 2009.

[16] D. Lopez-Perez, I. Guvenc, G. D. L. Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 22–30, 2011.

[17] G. Fettweis and E. Zimmermann, "ICT Energy Consumption - Trends and Challenges," in *Proc. WPMC*, LapLand, Finland, Sep 2008.

[18] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource Allocation and Cross-Layer Control in Wireless Networks," *Foundations and Trends® in Networking*, vol. 1, no. 1, 2006.

[19] M. J. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[20] L. Fleischer, "Data center scheduling, generalized flows, and submodularity," in *Proc. ANALCO*, Austin, USA, Jan 2010, pp. 56–65.

[21] U. Feige, V. Mirrokni, and J. Vondrak, "Maximizing non-monotone submodular functions," in *Proc. IEEE FOCS*, Rhode Island, USA, Jan 2007, pp. 461–471.

[22] N. Buchbinder, M. Feldman, J. S. Naor, and R. Schwartz, "A tight linear time (1/2)-approximation for unconstrained submodular maximization," in *Proc. IEEE FOCS*, New Brunswick, USA, Oct 2012, pp. 461–471.

[23] K. M. Anstreicher, "LINEAR PROGRAMMING IN $O(\frac{n^3}{\ln n}L)$ OPERATIONS," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 803–812, 1999.

[24] U. Feige, N. Immorlica, V. S. Mirrokni, and H. Nazerzadeh, "PASS Approximation A Framework for Analyzing and Designing Heuristics," in *In Proc. Approx + Random*, Aug 2009, pp. 111–124.

[25] S. Borst, M. Markakis, and I. Saniee, "Distributed power allocation and user assignment in ofdma cellular networks," in *In Proc. Allerton Conference on Communication, Control, and Computing*, Urbana, USA, Sep 2011, pp. 1055–1063.