Efficient Load Rearrangement of Small Cells with D2D Relay for Energy Saving and QoS Support

You-Chiun Wang  
Department of Computer Science and Engineering  
National Sun Yat-sen University  
Kaohsiung, Taiwan  
cywang@cse.nsysu.edu.tw

Zong-Han Lin  
Department of Computer Science and Engineering  
National Sun Yat-sen University  
Kaohsiung, Taiwan  
xiaopung83618@gmail.com

Abstract—To satisfy the growing demand of wireless access, a mass of small cells are deployed in the service area to intensify signal quality and team up with macrocells. However, it spends lots of energy to keep the operation of small cells, which collides with the goal of green communications. In the paper, we propose an efficient load sharing (ELS) scheme to conquer this problem by transferring services of user equipments (UEs) among different cells. For each small cell in the off-peak period, its serving UEs will be adaptively taken over by other cells through handover or user-to-network relay (i.e., D2D relay). Thus, its base station can switch to the sleep mode and save energy. The above mechanism is also applied to the small cells whose base stations are overloaded for mitigating congestion. Simulation results show that ELS raises energy efficiency, curtails energy expense of the base stations in small cells, and provides better QoS support for UEs.

Index Terms—device-to-device (D2D), discontinuous transmission (DTX), green communication, load sharing, small cell.

I. INTRODUCTION

Cell heterogeneity is an inevitable trend of development in mobile networks, where small cells are strategically deployed to enhance customer services [1], [2]. They not only help raise signal strength in some regions where crowds congregate like concert halls or malls, but also collaborate with macrocells for traffic-load sharing. A small-cell base station (BS) has many advantages over the traditional macrocell one in terms of price, transmitted power, and installation [3], [4].

Today, the global information communications technology (ICT) ecosystem spends towards 1800 Terawatt-hours of electricity per year, and the electricity demand continues to grow [5]. According to the estimation by the Mobile VCE [6], BSs consume more than 55% of total energy in mobile networks. It is imperative to support green communications to conserve energy and alleviate the greenhouse effect. However, deploying numerous BSs for improving network performance would be contrary to the objective of green communications.

In practice, the number of user equipments (UEs) served by a mobile network (and also their traffic demands) may fluctuate violently at different times [7]. For instance, there usually exist many UEs in a downtown office region during workdays but it becomes nearly empty in holidays. Moreover, a service area may be crowded with masses of UEs due to special activities such as singing concerts, whereas most of them often leave the area after activities finish. Evidently, it is uneconomic to keep operating small-cell BSs all the time.

To reduce wastage of electricity, the discontinuous transmission (DTX) technique [8] is commonly used to make idle BSs “sleep” by turning off their transceivers provisionally. However, DTX’s performance highly depends on the distribution of UEs [9]. Specifically, even though a small cell contains just a few UEs, its BS should keep active to serve them. Therefore, our aim is to “actively” make the BSs in off-peak periods (i.e., whose traffic loads are light) also go to sleep with the help of load rearrangement among small cells, so as to lift efficiency of DTX and achieve green communications.

In view of this, we propose an efficient load sharing (ELS) scheme to adaptively rearrange UEs served in some cells based on their traffic demands for energy saving and QoS support. For each small cell whose traffic load is not heavy, its UEs are efficiently handed over to neighboring cells to allow its BS going to sleep. For more flexibility, some of these UEs can connect with other BSs indirectly via user-to-network relay, which is realized by device-to-device (D2D) communications [10]. Besides, a overloaded BS can also ask other BSs to take over parts of the serving UEs to avoid congestion. In this way, the BS can have enough resources to serve UEs and meet their demands. Through simulations, we verify that ELS performs better than existing approaches in respect of energy efficiency (EE), energy expense of small-cell BSs, and also QoS support for UEs.

This paper is outlined as follows: Section II surveys related work and Section III elaborates on the ELS scheme. Then, performance evaluation is given in Section IV. Finally, Section V concludes this paper and gives future work.

II. RELATED WORK

Some DTX-related issues are addressed in the literature. To alleviate interference between cells, the study [11] shortens the amount of time in DTX when two nearby BSs wake up to transmit data. Saxena et al. [12] use the game theory to forecast traffic patterns of BSs using DTX. Sun et al. [13] exploit cell overlap to efficiently hand over the UEs served by a sleeping BS to others. The work [14] points out a ping-pong effect in DTX, where some BSs may be frequently switched on and off,
Fig. 1. A service area covered by macrocells and small cells.

and adds a hysteresis time to eliminate the effect. Evidently, these studies have different objectives with our paper.

Various strategies are also proposed to select small-cell BSs for sleeping in DTX. A non-cooperative game is adopted in [15] to let each BS decide whether to sleep, with the aim of minimizing a cost function. The work [16] chooses a subset of BSs to let each BS decide whether to sleep, with the aim of minimizing a cost function. The work [17] chooses a subset of BSs to keep the target throughput. In [17], a stochastic geometry tool is used to find a good density of BSs for service, and some BSs are randomly turned off based on the density. By modeling the locations of BSs as a Poisson point process, Li et al. [18] analyze the activation probability of BSs and also the coverage probability to enhance EE. The study [19] proposes a transmit-power scaling law (TPSL) to maintain network coverage, and derives the optimal ratio of deep-sleep BSs to maximize EE.

Only few studies consider actively transferring UEs among cells to let more BSs sleep. In [20], an energy-efficient pricing and resource scheduling (EPS) method is proposed to organize small cells into groups. A coordinator is picked in each group to transfer UEs among member cells for load balancing and also deactivate light-load BSs for energy saving. However, UE transfer across different groups is not allowed in EPS. Even though two small cells are neighbors, their UEs cannot be transferred if they are not in the same group. Our ELS scheme not only relaxes this limitation but also flexibly transfers UEs by user-to-network relay (via D2D communications), so it can save more energy of BSs and further improve EE.

III. THE PROPOSED ELS SCHEME

We are given a service area covered by LTE-A macrocells seamlessly. Small cells are placed inside macrocells to improve signal quality. There could exist crowded hotspot regions in the service area, where small cells are intensively deployed to enhance customer services. Fig. 1 shows an example.

Three modes are applied to manage BSs. A BS in the full-power (FP) mode offers services by the maximum transmitted power. In the sleep (SL) mode, the BS turns off its transceiver. Besides, a low-power (LP) mode is used when the BS serves more UEs but each UE’s demand is pretty small. In this case, the BS lowers its transmitted power to save energy. Macrocell BSs are always in the FP mode to provide seamless coverage. Our objective is to let more small-cell BSs sleep while meeting demands of more UEs, so as to support green communications. To do so, ELS first selects the mode of each BS, transfers UEs among cells, and then allot resources to UEs.

A. Mode Selection

To search a BS for service, each UE $u_i$ calculates its SINR with reference to each BS $b_j$ as follows:

$$\text{SINR}_{i,j} = \frac{P_{j,i}}{\varphi + \sum_{k \in \mathcal{E} \setminus b_j} P_{k,i}},$$

where $P_{j,i}$ is the amount of $b_j$’s power gotten by $u_i$, $\varphi$ is the power of white noise, and $\mathcal{E}$ is the set of BSs whose signals are captured by $u_i$. Then, SINR can be converted to a channel quality indicator (CQI) by Table I, which helps the BS pick a modulation and coding scheme (MCS) to send data. A default BS for $u_i$ is the one that provides the maximum CQI [22].

After that, the number $N_j$ of UEs served by BS $b_j$ is easily derived. To estimate the load $\hat{L}_j$ of $b_j$, we compute the number $m_i$ of resource blocks (RBs) required by each served UE $u_i$ based on its traffic demand $d_i$. Let $\lambda_i$ be the number of data bits carried by an RB (allocated to $u_i$), which depends on the MCS associated with the RB. Then, the least number of RBs required to meet $u_i$’s demand is calculated by

$$m_i^R = \arg \min_{m_i \in \mathbb{N}} \{\lambda_i m_i \geq r_i t\},$$

where $t$ is a transmission time interval (TTI = 1 ms). Given the number $M_j$ of RBs offered by $b_j$ in a TTI, its load is defined by $\hat{L}_j = \sum_{u_i \in U_j} m_i^R / M_j$, where $U_j$ denotes the set of UEs served by $b_j$.

By comparing $N_j$ and $\hat{L}_j$ with two thresholds $\zeta_N$ and $\zeta_L$, where $\zeta_N \in \mathbb{Z}^+$ and $0 < \zeta_L < 1$, we select the mode for each small-cell BS $b_j$ based on three cases. First, if $\hat{L}_j$ overtakes $\zeta_L$, which implies that $b_j$’s load is not light, $b_j$ should stay in the FP mode. Second, if both $\hat{L}_j < \zeta_L$ and $N_j < \zeta_N$, $b_j$ enters the off-peak period and can switch to the SL mode by UE transfer discussed in Section III-B. Third, when $\hat{L}_j < \zeta_L$ but $N_j \geq \zeta_N$, $b_j$ will change to the LP mode. Doing so has three advantages: 1) $b_j$ can save its energy on offering services; 2) when some UEs raise their demands suddenly, $b_j$ can fast switch to the FP mode without spending much energy to start

<table>
<thead>
<tr>
<th>CQI</th>
<th>MCS</th>
<th>SINR</th>
<th>CQI</th>
<th>MCS</th>
<th>SINR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QPSK (78)</td>
<td>-6.936</td>
<td>9</td>
<td>16QAM (616)</td>
<td>8.573</td>
</tr>
<tr>
<td>2</td>
<td>QPSK (120)</td>
<td>-5.147</td>
<td>10</td>
<td>64QAM (466)</td>
<td>10.366</td>
</tr>
<tr>
<td>3</td>
<td>QPSK (193)</td>
<td>-3.180</td>
<td>11</td>
<td>64QAM (567)</td>
<td>12.289</td>
</tr>
<tr>
<td>4</td>
<td>QPSK (308)</td>
<td>-1.253</td>
<td>12</td>
<td>64QAM (666)</td>
<td>14.173</td>
</tr>
<tr>
<td>5</td>
<td>QPSK (449)</td>
<td>2.699</td>
<td>13</td>
<td>64QAM (772)</td>
<td>15.888</td>
</tr>
<tr>
<td>6</td>
<td>QPSK (602)</td>
<td>7.100</td>
<td>14</td>
<td>64QAM (873)</td>
<td>17.814</td>
</tr>
<tr>
<td>7</td>
<td>16QAM (378)</td>
<td>4.694</td>
<td>15</td>
<td>64QAM (948)</td>
<td>19.829</td>
</tr>
<tr>
<td>8</td>
<td>16QAM (490)</td>
<td>6.525</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Each number in parentheses gives the code rate multiplied by 1024.*
up the transceiver; 3) it is easy for $b_j$ to accept and handle the UEs transferred from nearby cells. Theorem 1 then analyzes the time complexity of the mode selection method.

**Theorem 1.** Given $n_B$ BSs and $n_U$ UEs, the worst-case time complexity of the mode selection method is $O(n_Bn_U)$.

**Proof:** Computing SINRs of all UEs by Eq. (1) spends $O(n_Bn_U)$ time. Then, it takes $O(n_U)$ time to find the number of RBs used by each UE via Eq. (2). As each UE is served by a BS, finding the load of every BS takes $O(n_U)$ time. Finally, it consumes $O(n_B)$ time to decide the mode for each BS (by checking both $N_j$ and $L_j$). Thus, the overall time complexity is $O(n_Bn_U) + O(n_U) + O(n_U) + O(n_B) = O(n_Bn_U)$.

**B. UE Transfer**

To facilitate UE transfer, let us define four sets of small-cell BSs: 1) $E_{SL}$ is the set of BSs that will switch to the SL mode, 2) $E_{OL}$ is the set of overloaded BSs, whose $\hat{L}_j$ values exceed 1 (i.e., they do not have enough RBs to serve their UEs), 3) $E_{LP}$ is the set of BSs that can enter the LP mode, and 4) $E_{AL}$ is the set of BSs in the FP mode but their $L_j$ values are below a threshold $\delta$, where $\hat{L}_j < \delta < 1$ (e.g., $\delta = 0.8$). In other words, each BS in $E_{AL}$ still has unused RBs to help serve the UEs transferred from other cells.

For each BS $b_j$ in $E_{SL}$, we transfer all of its UEs to other cells, so as to make it sleep to save energy. Three rules are used to select a cell for UE transfer in sequence: R1) a small cell whose BS is in $E_{SL}$, R2) a small cell whose BS is in $E_{LP}$, and R3) the macrocell where $b_k$ locates. Specifically, our idea is to first select those small-cell BSs already in the FP mode to serve $b_j$’s UEs (i.e., rule R1), as they need not spend more energy to serve these extra UEs. However, once the load $\hat{L}_k$ of such a BS $b_k$ overtakes threshold $\delta$, $b_k$ will not accept extra UEs and be removed from $E_{AL}$. The reason is that $b_k$ should reserve a few unused RBs to handle the situation where some of its serving UEs raise demands suddenly. Then, rule R2 is adopted when $b_j$ cannot find any BS in $E_{SL}$ to serve its UEs. In this case, $b_j$ asks a BS $b_k$ in the LP mode to serve its UEs. Note that if $\hat{L}_k \geq \zeta$ after serving the extra UEs, $b_k$ has to switch to the FP mode (and be moved from set $E_{LP}$ to set $E_{AL}$). However, if no BS can be found from both sets $E_{AL}$ and $E_{LP}$, rule R3 is applied. Thus, $b_j$’s UEs will be transferred to the macrocell to let $b_j$ switch to the SL mode.

We employ two strategies to transfer $b_j$’s UEs to a selected cell: 1) user-to-network relay (also called D2D relay) [23] and 2) traditional handover [24]. Strategy 1 has a higher priority than strategy 2 to provide flexibility. Fig. 2 shows how to select D2D relay in strategy 1. Suppose that UE $u_i$ is served by BS $b_j$ and UE $u_x$ is served by BS $b_y$. Then, $u_i$ can be transferred to $b_y$’s cell via $u_x$’s D2D relay if $b_y$ has enough resources to accept the D2D relay and two conditions are also met:

$$\frac{(\text{SINR}_{i,x} + \text{SINR}_{y,y})}{2} \geq \text{SINR}_{i,j},$$

$$\min\{\text{SINR}_{i,x}, \text{SINR}_{y,y}\} \geq \sigma \text{SINR}_{i,j},$$

where $\sigma$ is a ratio close to one (e.g., $\sigma = 0.9$). Here, Eq. (3) indicates that the average signal quality of links $(u_i, u_x)$ and $(u_x, b_y)$ should be at least as good as that of $u_i$’s original link $(u_i, b_j)$. Eq. (4) is to avoid the extreme case where either of links $(u_i, u_x)$ and $(u_x, b_y)$ has pretty bad signal quality but Eq. (3) still holds. Both Eqs. (3) and (4) make sure that the service quality of $u_i$ will not significantly degrade (or even can raise) after it is transferred to $b_y$’s cell by D2D relay.

On the other hand, for each BS $b_j$ in $E_{OL}$, some of its UEs are also transferred to other cells to ensure that $b_j$ has enough RBs to support QoS for residual UEs. Let $U_j$ be the set of UEs served by $b_j$. We iteratively pick a UE $u_i$ from $U_j$ with the minimum CQI, and transfer it to a nearby cell based on the three rules. However, to avoid congesting the macrocell, rule R3 will not be used if the macrocell BS becomes overloaded. This iteration is repeated until $\hat{L}_j \leq \delta$ or no UEs in $U_j$ can be transferred. After that, $b_j$ is removed from set $E_{OL}$. Theorem 2 gives the time complexity of the UE transfer method.

**Theorem 2.** The UE transfer method spends time of $(n_B^{SL} + n_B^{OL}) \times (n_B^{AL} + n_B^{LP} + n_B^{M}) + (\rho + 1) \times (n_U^{SL} + n_U^{OL} - n_B^{OL})$, where $n_B^{SL}$, $n_B^{OL}$, $n_B^{AL}$, and $n_B^{LP}$ are the numbers of BSs in sets $E_{SL}$, $E_{OL}$, $E_{AL}$, and $E_{LP}$, respectively, $n_B^{M}$ is the number of macrocell BSs, $\rho$ is the average number of neighbors of each UE, and $n_U^{SL}$ and $n_U^{OL}$ are the numbers of UEs served by the BSs in sets $E_{SL}$ and $E_{OL}$, respectively.

**Proof:** For each BS in $E_{SL}$, we employ the three rules to check if other BSs can take over its UEs. The worst case is to check all small-cell BSs in both sets $E_{AL}$ and $E_{LP}$ and also macrocell BSs. It thus takes time of $n_B^{SL} \times (n_B^{AL} + n_B^{LP} + n_B^{M})$. Then, we adopt D2D relay and handover to transfer each UE served by BSs in $E_{OL}$, which will spend time of $n_U^{OL} \times \rho$ and $n_U^{SL}$, respectively. After that, we check if other BSs can take over parts of UEs served by BSs in $E_{OL}$, which takes time of $n_B^{OL} \times (n_B^{AL} + n_B^{LP} + n_B^{M})$. Since at least one UE should be served by each BS in $E_{OL}$ (or the BS will be idle), using D2D relay and handover to transfer parts of UEs served by BSs in $E_{OL}$ spends time of $(n_U^{OL} - n_B^{OL}) \times \rho$ and $n_U^{OL} - n_B^{OL}$, respectively. By taking the sum of the above time and doing some algebraic operations, we can thus verify this theorem.

**C. Resource Allocation**

For a BS $b_j$ with $\hat{L}_j \leq 1$, we can simply allocate a number $n_U^R$ of RBs in Eq. (2) to each of its serving UE to support QoS. However, if the BS does not have enough RBs to meet demands of all UEs (i.e., $\hat{L}_j > 1$), we allocate RBs to each UE based on the tax mechanism in [25] (with some modifications).

Consider that $b_j$ serves a set $U_j$ of UEs. For each RB offered by $b_j$, it is allocated to a UE in $U_j$ with the maximum CQI in terms of that RB. Let $U_j^{urg}$ be the subset of urgent UEs in $U_j$, where their packets are about to be dropped in the next TTI.
For each UE $u_i$ in $U_j - U_j^{urg}$, since it is not in danger of packet expiration, we can ask $u_i$ to return a part of acquiring RBs (i.e., tax) to relieve urgent UEs. In particular, if $u_i$ acquires $n_i$ RBs, where $n_i > \alpha$, it will be taxed with a number $[\beta (m_i - \alpha)]$ of RBs, where $\alpha \in \mathbb{Z}^+$ and $0 < \beta \leq 1$. For example, suppose that three UEs $u_1$, $u_2$, and $u_3$ in $U_j - U_j^{urg}$ acquire 2, 4, and 6 RBs, respectively. If we set $\alpha = 2$ and $\beta = 0.3$, then $u_1$, $u_2$, and $u_3$ have to return 0, 1, and 2 RBs, respectively.

After that, the taxed RBs are distributed among urgent UEs. We tailor the tax mechanism in [25] to our needs. Specifically, all UEs in $U_j^{urg}$ are sorted based on their amount of expired data in the next TTI decreasingly. Then, we allocate one taxed RB to each urgent UE in a round-robin manner, until either all taxed RBs are used up or no UEs in $U_j^{urg}$ will encounter packet loss in the next TTI (due to getting enough taxed RBs). For the latter case, we can give back the remaining taxed RBs to their original owners. Theorem 3 discusses the time complexity of the resource allocation method.

Theorem 3. Suppose that a BS offers $n_R$ RBs in a TTI, and it serves $n_j$ UEs (with $n_j^{urg}$ urgent UEs). The resource allocation method then takes time of $O(n_R n_j + n_j^{urg} \log n_j^{urg})$.

Proof: Each RB is first given to a UE with the maximum CQI. Since there are $n_R$ RBs and $n_j$ UEs, this operation takes $O(n_R n_j)$ time. Then, each UE in $U_j - U_j^{urg}$ is taxed with some RBs, which spends $O(n_j - n_j^{urg})$ time. Finally, urgent UEs are sorted to get taxed RBs by round robin, which consumes time of $O(n_j^{urg} \log n_j^{urg} + n_j^{urg})$. To sum up, the time complexity is $O(n_R n_j) + O(n_j - n_j^{urg}) + O(n_j^{urg} \log n_j^{urg} + n_j^{urg}) = O(n_R n_j + n_j^{urg} \log n_j^{urg})$.

IV. PERFORMANCE EVALUATION

We use MATLAB to evaluate performance. Fig. 1 gives the network topology with 7 macrocells and 64 small cells. About 40% small cells lie in the hotspot region. Other small cells are placed near the boundary of each macrocell to tone up signals. A macrocell has radius of 1500m. Its BS has transmitted power of 46dBm and channel bandwidth of 20MHz, which offers 100 RBs every TTI. The radius of a small cell is 250m. Its BS has transmitted power of 30dBm and channel bandwidth of 5MHz, which provides 25 RBs in a TTI.

For wireless transmissions, we employ a log-distance model to measure the amount of signal’s attenuation caused by path loss from a BS $b_j$ to a serving UE $u_i$: $128.1 + 37.6 \log d(u_i, b_j)$ for macrocells and $38 + 30 \log(10^3 d(u_i, b_j))$ for small cells, where $d(u_i, b_j)$ is their distance in kilometers [4]. Moreover, a zero-mean log-normal distribution is used to estimate the effect of shadowing fading. Its standard deviation is set to 10dB and 6dB for macrocells and small cells, respectively. The power spectral density of the white noise is set to -174dBm/Hz.

There are 300 to 1800 UEs in the service area, where 3/4 of them congregate in the hotspot region to simulate crowds. Each UE produces one of the following flows: 1) 8.4kbps VoIP flow, 2) 242kbps H.264 video flow, and 3) 12kbps non-real-time flow. The delay budget of VoIP and video flows is 100ms. Two scenarios are considered. In scenario A, the numbers of VoIP, video, and non-real-time UEs are equal. In scenario B, we set the ratio of VoIP, video, and non-real-time UEs to 2:2:1. Therefore, the total traffic demand in scenario B will be much larger than that in scenario A, as there are more video UEs.

We compare our ELS scheme with two methods discussed in Section II, including TPSL [19] and EPS [20]. In particular, TPSL finds an optimal ratio of deep-sleep BSs to increase EE. EPS transfers UEs to let more small-cell BSs sleep based on a grouping policy. If a small-cell BS is in the FP, LP, and SL modes, it spends 100%, 50%, and 10% of energy for operation [26]. Moreover, we set $\zeta_L$ to 0.3 and $\zeta_N$ to 1/3 of the average number of UEs in each small cell.

A. Energy Efficiency (EE)

We evaluate the amount of EE, which is defined by the ratio of network throughput to energy expense of BSs (measured in kb/W) [27]. Higher EE implies that BSs could better utilize their transmitted power to send more user data. Fig. 3(a) and Fig. 3(b) give the amount of EE in scenarios A and B, respectively. Because the total traffic demand grows substantially as there are more UEs, network throughput will increase, which raises EE. By comparing Fig. 3(a) with Fig. 3(b), we can also get a similar observation, since the traffic demand in scenario B is larger than that in scenario A.

Comparing with TPSL, EPS actively transfers UEs to help more small-cell BSs sleep, so EPS has higher EE than TPSL. Our proposed ELS scheme not only transfers UEs more efficiently in contrast to EPS (by relaxing the grouping limitation and using D2D relay) but also improves the performance of resource allocation by the modified tax mechanism. Thus, ELS results in the highest EE, which also reveals its effectiveness on supporting green communications.

B. Energy Expense of Small-cell BSs

Since macrocell BSs are always in the FP mode for every method, we measure the amount of energy consumed by small-cell BSs in both scenarios, as shown in Fig. 3(c) and Fig. 3(d).

Here, the reduction ratio gives the percentage of energy further saved by ELS as compared with TPSL and EPS. In general, small-cell BSs spend more energy when there are more UEs or in scenario B. The reason is that more small-cell BSs have to stay in the FP mode and use the maximum power to satisfy the growing demands of user traffics. Our ELS scheme not only transfers small-cell BSs enter the SL mode by adaptively transferring UEs but also takes advantage of the LP mode to let some non-sleeping BSs cut down their energy expense to serve UEs with lower traffic demands. Therefore, ELS can save more energy of small-cell BSs than both TPSL and EPS, especially when there are fewer UEs. Even though there are 1800 UEs in the service area, ELS still saves around 18.9% and 13.5% of energy as compared with TPSL and EPS, respectively.

C. QoS Support for UEs

To assess QoS support, we measure the packet loss rate of real-time flows (i.e., VoIP and video flows) and the number of UEs whose traffic demands are satisfied by each method.
Fig. 3(e) and Fig. 3(f) present the experimental data in scenarios A and B, respectively. When the number of UEs grows, the packet loss rate also increases, because more UEs compete for the fixed resources. Such a phenomenon is especially obvious in scenario B, as there are more bandwidth-consuming video flows. On the other hand, the curve of satisfied UEs ascends gradually to a plateau at 1200 UEs. The reason is that the total number of available RBs is constant, and the network becomes nearly saturate when there are more than 1200 UEs.

From the result in Fig. 3(e) and Fig. 3(f), our ELS scheme always has the lowest packet loss rate and satisfies the most number of UEs among all methods. Thus, ELS can offer better QoS support for UEs, as compared with TPSL and EPS.

D. Effect of Threshold $\zeta_L$

Recall that threshold $\zeta_L$ is used to judge whether a small-cell BS should switch to the FP mode (i.e., using the maximum transmitted power to provide services). In this experiment, we thus study its effect on ELS’s performance in both scenarios, as shown in Fig. 3(g) and Fig. 3(h). Generally speaking, fewer small-cell BSs will switch to the FP mode when $\zeta_L$ increases (referring to the mode selection in Section III-A). In this case, the packet loss rate will increase whereas the amount of energy expense of small-cell BSs can decrease (because more of them can enter the LP or SL modes to save energy). Based on the result in Fig. 3(g) and Fig. 3(h), we suggest setting $\zeta_L$ to 0.3, so as to save more energy of small-cell BSs while keeping a lower packet loss rate.

E. Daily Energy Consumption of BSs

To investigate how each method reacts to the change of the number of UEs as time goes by, we use the daily traffic profile of Europe [28] to simulate traffic demands in one day, where the maximum number of UEs is set to 1800. Fig. 3(i) presents the overall energy consumption of all BSs in scenario A. It can be observed that the off-peak period is between 2:00 and 9:00. By actively transferring UEs among cells, both EPS and ELS can reduce more energy consumption of BSs than TPSL, even
in the peak period. Our ELS scheme allows UEs to be freely transferred (i.e., without the grouping restriction in EPS) and employs D2D relay to flexibly transfer UEs, so it can further reduce energy consumption as compared with EPS, especially in the off-peak period. This experiment shows that ELS better supports green communications than other methods.

V. CONCLUSION AND FUTURE WORK

Deploying many small cells in a mobile network has become the inexorable trend, but keeping operating all small-cell BSs consumes too much energy and violates the objective of green communications. To conquer this dilemma, we propose the ELS scheme to efficiently transfer UEs among cells by D2D relay and handover. Not only more small-cell BSs that serve just few UEs are allowed to sleep to conserve energy, but also overloaded BSs can share out their loads with others to support QoS for more UEs. Through MATLAB simulations, we verify that ELS can significantly improve EE, reduce energy expense of small-cell BSs, and provide better QoS support for UEs, as compared with the existing TPSL and EPS methods.

By simulations, we show that both packet loss and energy expense could be balanced by setting threshold $\zeta_L$ to 0.3 in the ELS scheme. For the future work, we will analyze the optimal value of $\zeta_L$ based on different parameters. Furthermore, it is interesting to consider the effect of UE mobility on ELS [29]. Finally, how to provide transmission fairness among UEs in different cells deserves further investigation [30].

ACKNOWLEDGMENT

You-Chiun Wang’s research is co-sponsored by the Ministry of Science and Technology under Grant No. MOST 108-2221-E-110-016-MY3, Taiwan.

REFERENCES