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
ycwang@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 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,
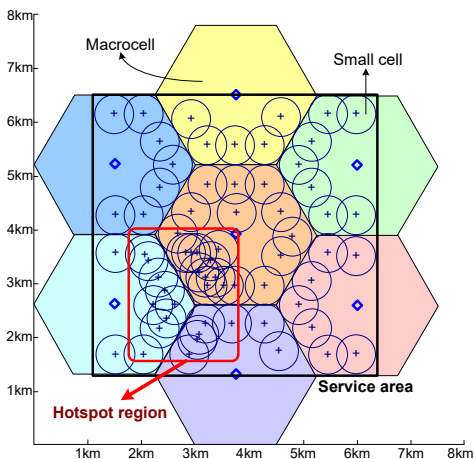
Fig. 1. A service area covered by macrocells and small cells.

CQI	MCS	SINR	CQI	MCS	SINR
1	QPSK (78)†	-6.936	9	16QAM (616)	8.573
2	QPSK (120)	-5.147	10	64QAM (466)	10.366
3	QPSK (193)	-3.180	11	64QAM (567)	12.289
4	QPSK (308)	-1.253	12	64QAM (666)	14.173
5	QPSK (449)	0.761	13	64QAM (772)	15.888
6	QPSK (602)	2.699	14	64QAM (873)	17.814
7	16QAM (378)	4.694	15	64QAM (948)	19.829
8	16QAM (490)	6.525	⋄ SINR is measured in dB.		

†Each number in parentheses gives the code rate multiplied by 1024.

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 allots 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:

$$\text{SINR}_{i,j} = \frac{P_{j,i}}{\varphi + \sum_{b_k \in \mathcal{E}, b_k \neq b_j} P_{k,i}}, \quad (1)$$

where $P_{j,i}$ is the amount of b_j's power gotten by u_i, φ 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 $\hat{N}_j$ 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_j. Let λ_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

$$m_i^{\mathbf{R}} = \arg\min_{m_i}\{\lambda_i m_i \geq r_i t\}, \quad (2)$$

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 $\hat{L}_j = \sum_{u_i \in U_j} m_i^{\mathbf{R}}/M_j$, where U_j denotes the set of UEs served by b_j.

By comparing $\hat{N}_j$ and $\hat{L}_j$ with two thresholds ζ_N and ζ_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 ζ_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 $\hat{N}_j < \zeta_N$, 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 $\hat{N}_j \geq \zeta_N$, 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

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 enter the sleep mode based on their traffic loads, so as 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
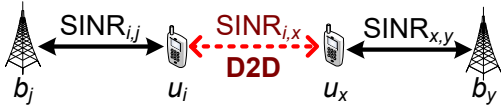
Fig. 2. Selection of D2D relay.

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 $O(n_B n_U)$.*

Proof: Computing SINRs of all UEs by Eq. (1) spends $O(n_B n_U)$ 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 $\hat{N}_j$ and $\hat{L}_j$). Thus, the overall time complexity is $O(n_B n_U) + O(n_U) + O(n_U) + O(n_B) = O(n_B n_U)$. ∎

B. UE Transfer

To facilitate UE transfer, let us define four sets of small-cell BSs: 1) $\mathcal{E}_{SL}$ is the set of BSs that will switch to the SL mode, 2) $\mathcal{E}_{OL}$ 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) $\mathcal{E}_{LP}$ is the set of BSs that can enter the LP mode, and 4) $\mathcal{E}_{AL}$ is the set of BSs in the FP mode but their $\hat{L}_j$ values are below a threshold δ, where $\zeta_L < \delta < 1$ (e.g., $\delta = 0.8$). In other words, each BS in $\mathcal{E}_{AL}$ still has unused RBs to help serve the UEs transferred from other cells.

For each BS b_j in $\mathcal{E}_{SL}$, 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 $\mathcal{E}_{AL}$, **R2)** a small cell whose BS is in $\mathcal{E}_{LP}$, and **R3)** the macrocell where b_j 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 δ, b_k will not accept extra UEs and be removed from $\mathcal{E}_{AL}$. 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 $\mathcal{E}_{AL}$ to serve its UEs. In this case, b_j asks a BS b_h in the LP mode to serve its UEs. Note that if $\hat{L}_h \geq \zeta_L$ after serving the extra UEs, b_h has to switch to the FP mode (and be moved from set $\mathcal{E}_{LP}$ to set $\mathcal{E}_{AL}$). However, if no BS can be found from both sets $\mathcal{E}_{AL}$ and $\mathcal{E}_{LP}$, 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:

$$(\text{SINR}_{i,x} + \text{SINR}_{x,y})/2 \geq \text{SINR}_{i,j}, \qquad (3)$$

$$\min\{\text{SINR}_{i,x}, \text{SINR}_{x,y}\} \geq \sigma \text{SINR}_{i,j}, \qquad (4)$$

where σ 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 $\mathcal{E}_{OL}$, 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 $\mathcal{E}_{OL}$. 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 $\mathcal{E}_{SL}$, $\mathcal{E}_{OL}$, $\mathcal{E}_{AL}$, and $\mathcal{E}_{LP}$, respectively, n_B^{M} is the number of macrocell BSs, ρ 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 $\mathcal{E}_{SL}$ and $\mathcal{E}_{OL}$, respectively.*

Proof: For each BS in $\mathcal{E}_{SL}$, 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 $\mathcal{E}_{AL}$ and $\mathcal{E}_{LP}$ 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 $\mathcal{E}_{SL}$, which will spend time of $n_U^{SL} \times \rho$ and n_U^{SL}, respectively. After that, we check if other BSs can take over parts of UEs served by BSs in $\mathcal{E}_{OL}$, 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 $\mathcal{E}_{OL}$ (or the BS will be idle), using D2D relay and handover to transfer parts of UEs served by BSs in $\mathcal{E}_{OL}$ 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 m_i^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 $U_j - U_j^{\mathrm{urg}}$, 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 m_i RBs, where $m_i > \alpha$, it will be taxed with a number $\lceil \beta \times (m_i - \alpha) \rceil$ of RBs, where $\alpha \in \mathbb{Z}^+$ and $0 < \beta \le 1$. For example, suppose that three UEs u_1, u_2, and u_3 in $U_j - U_j^{\mathrm{urg}}$ 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 $O(n_{\mathrm{R}} n_j + n_j^{\mathrm{urg}} \lg n_j^{\mathrm{urg}})$.*

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 $O(n_{\mathrm{R}} n_j)$ time. Then, each UE in $U_j - U_j^{\mathrm{urg}}$ is taxed with some RBs, which spends $O(n_j - n_j^{\mathrm{urg}})$ time. Finally, urgent UEs are sorted to get taxed RBs by round robin, which consumes time of $O(n_j^{\mathrm{urg}} \lg n_j^{\mathrm{urg}} + n_j^{\mathrm{urg}})$. To sum up, the time complexity is $O(n_{\mathrm{R}} n_j) + O(n_j - n_j^{\mathrm{urg}}) + O(n_j^{\mathrm{urg}} \lg n_j^{\mathrm{urg}} + n_j^{\mathrm{urg}}) = O(n_{\mathrm{R}} n_j + n_j^{\mathrm{urg}} \lg n_j^{\mathrm{urg}})$. ∎

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 ζ_L to 0.3 and ζ_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 helps more 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.
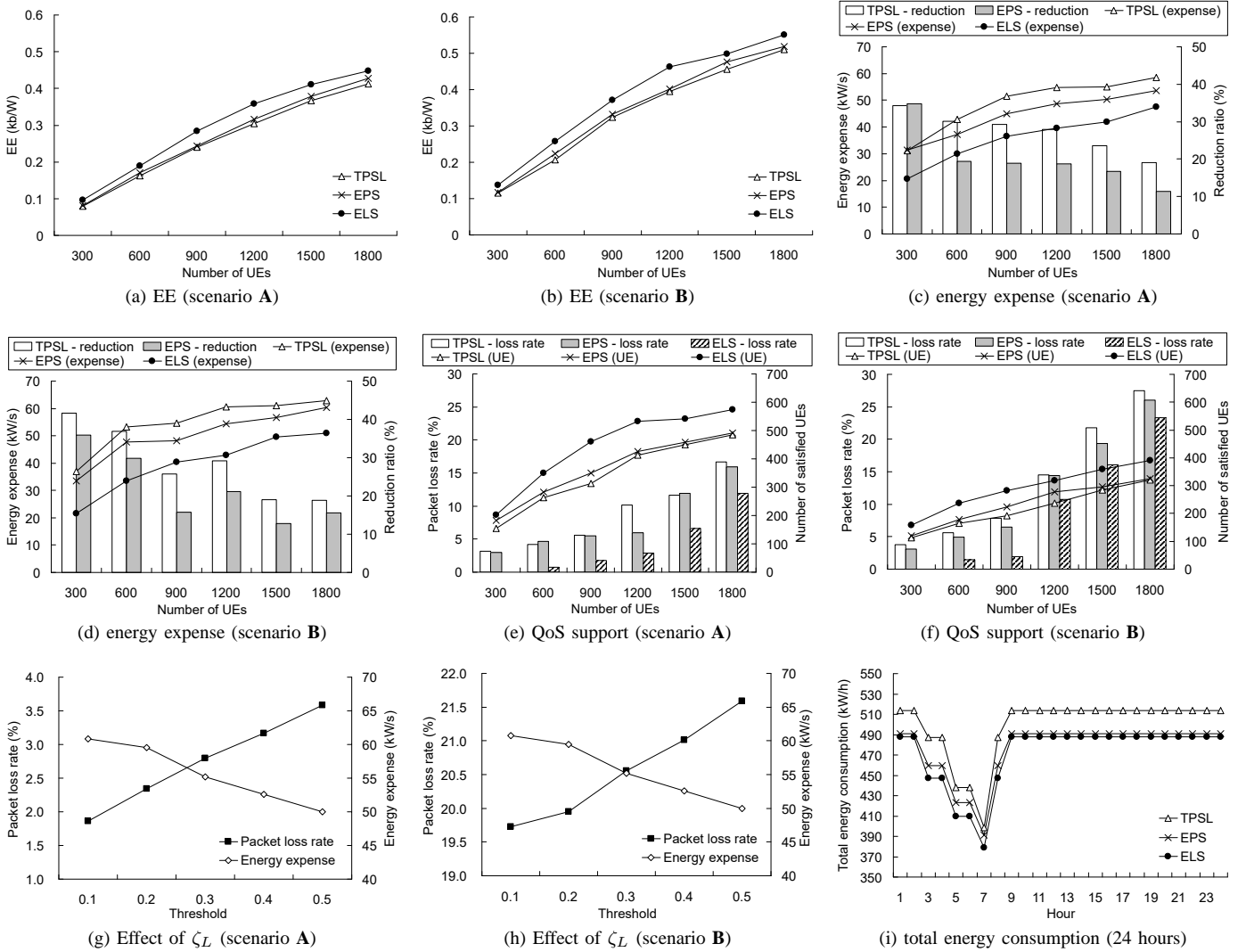
(a) EE (scenario **A**)   (b) EE (scenario **B**)   (c) energy expense (scenario **A**)

(d) energy expense (scenario **B**)   (e) QoS support (scenario **A**)   (f) QoS support (scenario **B**)

(g) Effect of ζ_L (scenario **A**)   (h) Effect of ζ_L (scenario **B**)   (i) total energy consumption (24 hours)

Fig. 3. Experimental results.

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 ζ_L

Recall that threshold ζ_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 ζ_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 ζ_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 ζ_L to 0.3 in the ELS scheme. For the future work, we will analyze the optimal value of ζ_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].

References

[1] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: challenges and solution approaches," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 22–31, 2015.

[2] Y. C. Wang and S. T. Chen, "Delay-aware ABS adjustment to support QoS for real-time traffic in LTE-A HetNet," *IEEE Wireless Communications Letters*, vol. 6, no. 5, pp. 590–593, 2017.

[3] H. Zhang, Y. Dong, J. Cheng, M. J. Hossain, and V. C. M. Leung, "Fronthauling for 5G LTE-U ultra dense cloud small cell networks," *IEEE Wireless Communications*, vol. 23, no. 6, pp. 48–53, 2016.

[4] Y. C. Wang and S. Lee, "Small-cell planning in LTE HetNet to improve energy efficiency," *International Journal of Communication Systems*, vol. 31, no. 5, pp. 1–18, 2018.

[5] R. Mahapatra, Y. Nijsure, G. Kaddoum, N. U. Hassan, and C. Yuen, "Energy efficiency tradeoff mechanism towards wireless green communication: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 686–705, 2016.

[6] Mobile VCE, "Green radio," http://www.mobilevce.com/green-radio.

[7] Y. C. Wang and C. A. Chuang, "Efficient eNB deployment strategy for heterogeneous cells in 4G LTE systems," *Computer Networks*, vol. 79, pp. 297–312, 2015.

[8] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in LTE with cell DTX," in *IEEE Vehicular Technology Conference*, 2011, pp. 1–5.

[9] Y. C. Wang and K. C. Chien, "A load-aware small-cell management mechanism to support green communications in 5G networks," in *IEEE Wireless and Optical Communication Conference*, 2018, pp. 1–5.

[10] W. K. Lai, Y. C. Wang, H. C. Lin, and J. W. Li, "Efficient resource allocation and power control for LTE-A D2D communication with pure D2D model," *IEEE Transactions on Vehicular Technology*, 2020.

[11] K. Abdallah, I. Cerutti, and P. Castoldi, "Energy-efficient coordinated sleep of LTE cells," in *IEEE International Conference on Communications*, 2012, pp. 5238–5242.

[12] N. Saxena, B. J. R. Sahu, and Y. S. Han, "Traffic-aware energy optimization in green LTE cellular systems," *IEEE Communications Letters*, vol. 18, no. 1, pp. 38–41, 2014.

[13] Y. Sun, Y. Chang, S. Song, and D. Yang, "An energy-efficiency aware sleeping strategy for dense multi-tier HetNets," in *IEEE Globecom*, 2014, pp. 1180–1185.

[14] R. Combes, S. E. Elayoubi, A. Ali, L. Saker, and T. Chahed, "Optimal online control for sleep mode in green base stations," *Computer Networks*, vol. 78, no. 26, pp. 140–151, 2015.

[15] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Opportunistic sleep mode strategies in wireless small cell networks," in *IEEE International Conference on Communications*, 2014, pp. 2707–2712.

[16] Y. L. Chung, "An efficient power-saving transmission mechanism in LTE macrocell-femtocell hybrid networks," in *International Conference on Information Networking*, 2014, pp. 176–180.

[17] Z. Kailai, L. Tiejun, and G. Hui, "A stochastic geometry based two-stage energy consumption minimization strategy via sleep mode with QoS constraint," in *IEEE International Conference on Communications*, 2016, pp. 1–6.

[18] L. Li, M. Peng, C. Yang, and Y. Wu, "Optimization of base-station density for high energy-efficient cellular networks with sleeping strategies," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7501–7514, 2016.

[19] P. Chang and G. Miao, "Optimal operation of base stations with deep sleep and discontinuous transmission," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 113–11 126, 2018.

[20] Y. C. Wang and K. C. Chien, "EPS: energy-efficient pricing and resource scheduling in LTE-A heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8832–8845, 2018.

[21] C. Mehlfuhrer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp, "Simulating the long term evolution physical layer," in *European Signal Processing Conference*, 2009, pp. 1471–1478.

[22] Y. C. Wang and T. Y. Tsai, "A pricing-aware resource scheduling framework for LTE networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1445–1458, 2017.

[23] R. Ma, Y. J. Chang, H. H. Chen, and C. Y. Chiu, "On relay selection schemes for relay-assisted D2D communications in LTE-A systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8303–8314, 2017.

[24] D. Castro-Hernandez and R. Paranjape, "Optimization of handover parameters for LTE/LTE-A in-building systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 5260–5273, 2018.

[25] Y. C. Wang and S. Y. Hsieh, "Service-differentiated downlink flow scheduling to support QoS in long term evolution," *Computer Networks*, vol. 94, pp. 344–359, 2016.

[26] E. Mugume and D. K. C. So, "Sleep mode mechanisms in dense small cell networks," in *IEEE International Conference on Communications*, 2015, pp. 192–197.

[27] Y. C. Wang and C. C. Huang, "Efficient management of interference and power by jointly configuring ABS and DRX in LTE-A HetNets," *Computer Networks*, vol. 150, pp. 15–27, 2019.

[28] G. Auer, O. Blume, and V. Giannini, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," EARTH Project Report, Deliverable D2.3, 2012.

[29] W. H. Yang, Y. C. Wang, Y. C. Tseng, and B. S. P. Lin, "Energy-efficient network selection with mobility pattern awareness in an integrated WiMAX and WiFi network," *International Journal on Communication Systems*, vol. 23, no. 2, pp. 213–230, 2010.

[30] Y. C. Wang and D. R. Jhong, "Efficient allocation of LTE downlink spectral resource to improve fairness and throughput," *International Journal of Communication Systems*, vol. 30, no. 14, pp. 1–13, 2017.