This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2850843, IEEE Transactions on Green Communications and Networking

Optimal Dual-Connectivity Traffic Offloading in Energy-Harvesting Small-Cell Networks

Yuan Wu *Senior Member IEEE*, Xiaowei Yang, Li Ping Qian *Senior Member IEEE*, Haibo Zhou *Member IEEE*, Xuemin (Sherman) Shen *Fellow IEEE*, Mohamad Awad *Member IEEE*

***Abstract*—Traffic offloading through heterogenous small-cell networks (HSCNs) has been envisioned as a cost-efficient approach to accommodate the tremendous traffic growth in cellular networks. In this paper, we investigate an energy-efficient dual-connectivity (DC) enabled traffic offloading through HSCNs, in which small cells are powered in a hybrid manner including both the conventional on-grid power-supply and renewable energy harvested from environment. To achieve a flexible traffic offloading, the emerging DC-enabled traffic offloading in 3GPP specification allows each mobile user (MU) to simultaneously communicate with a macro cell and offload data through a small cell. In spite of saving the on-grid power consumption, powering traffic offloading by energy harvesting (EH) might lead to quality of service (QoS) degradation, e.g., when the EH power-supply fails to support the required offloading rate. Thus, to reap the benefits of the DC-capability and the EH power-supply, we propose a joint optimization of traffic scheduling and power allocation that aims at minimizing the total on-grid power consumption of macro and small cells, while guaranteeing each served MU's traffic requirement. We start by studying a representative case of one small cell serving a group of MUs. In spite of the non-convexity of the formulated joint optimization problem, we exploit its layered structure and propose an algorithm that efficiently computes the optimal offloading solution. We further study the scenario of multiple small cells, and investigate how the small cells select different MUs for maximizing the system-wise reward that accounts for the revenue for offloading the MUs' traffic and the cost of total on-grid power consumption of all cells. We also propose an efficient algorithm to find the optimal MU-selection solution. Numerical results are provided to validate our proposed algorithms and show the advantage of our proposed DC-enabled traffic offloading through the EH-powered small cells.**

I. Introduction

The past decade has witnessed an explosive growth of smart mobile devices and popularity of mobile internet services, which have yielded a tremendous traffic burden on cellular networks. By exploiting the multi-tier structure of radio access networks (RANs), offloading mobile traffic through heterogenous small-cell networks (HSCNs) has been widely considered as a cost-efficient approach to relieve traffic congestion in macro cells. Due to bringing RANs closer to mobile users (MUs), traffic offloading through HSCNs can yield multi-fold benefits, such as enhancing throughput and improving resource utilization efficiency. To facilitate a flexible traffic offloading, the recent 3GPP specification has proposed a paradigm of small-cell dual-connectivity (DC) that enables a MU, by using two different radio interfaces, to communicate with a macro cell and simultaneously offload data through small cells [1]. With a flexible traffic scheduling between macro and small cells, the DC-enabled traffic scheduling is expected to further enhance the benefit of traffic offloading [2]–[7].

However, the dense deployment of HSCNs has yielded a significant energy consumption in cellular RANs, which has attracted lots of attentions in realizing energy-efficient HSCNs while providing guaranteed quality of service (QoS) [9]–[11]. Viewing the important role of traffic offloading, many research efforts have also been devoted to investigating energy-efficient traffic offloading through HSCNs [12]. In particular, with the recent advances in collecting and storing renewable energy from environment (e.g., via the emerging smart grids [20]), traffic offloading through small cells which are powered by energy-harvesting (EH) has been considered as a viable approach to reap the benefit of traffic offloading and to reduce the on-grid power consumption [21]–[25].

However, due to the randomness in renewable energy sources, the EH power-supply suffers from intermittency, which adversely influences the performance of traffic offloading. For instance, severe offloading outage (e.g., packet loss) due to insufficient received signal to noise and interference ratio (SINR) will occur if the small cell blindly provides a large offloading rate while suffering from a temporary valley of the EH power-supply. Several studies have been denoted to investigating how to properly exploit the intermittent EH power-supply to accommodate the MUs' traffic. In particular, the emerging paradigm of DC, which allows the MU to simultaneously communicate with the macro and small cells, enables a flexible traffic scheduling between macro and small cells [1], [2] and thus provides an effective approach to address the intermittent EH power-supply. For instance, based on the DC, the small cell suffering from the valley of EH power-supply can slow down its offloading rate to the MU, and correspondingly, the macro cell actively increases its transmission rate to the MU in order to maintain

Y. Wu, L. Qian, and X. Yang are with College of Information Engineering, Zhejiang University of Technology, Hangzhou, China (emails: iewuy@zjut.edu.cn, lpqian@zjut.edu.cn). L. Qian is the corresponding author. Y. Wu is also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xian, 710071, China.

H. Zhou and X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (email: h53zhou@uwaterloo.ca, xshen@bbcr.uwaterloo.ca). H. Zhou is now with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China.

M. Awad is with the Department of Computer Engineering, Kuwait University, 13060 Kuwait City, Kuwait (email: mohamad@ieee.org).

This work has been presented, in part, in IEEE Global Communications Conference (GLOBECOM) 2017.

This work was supported in part by the National Natural Science Foundation of China under Grant 61572440, in part by the Zhejiang Provincial Natural Science Foundation of China under Grants LR17F010002 and LR16F010003, in part by the Young Talent Cultivation Project of Zhejiang Association for Science and Technology under Grant 2016YCGC011, and in part by the Natural Sciences and Engineering Research Council, Canada.

the required QoS (e.g., throughput). Therefore, in this study, by exploiting the EH power-supply and the DC-capability, we investigate the DC-enabled traffic offloading through the EH-powered small cells. The main contributions of this paper are summarized as follows.

- We start by studying the scenario of one EH-powered small-cell access point (sAP) which offloads traffic for a group of MUs. Specifically, given the total number of served MUs, we formulate a joint optimization of the traffic scheduling and power allocation for one targeted pair of the sAP and the MU. Our formulation takes into account the offloading outage due to the sAP's intermittent EH power-supply and aims at minimizing the total on-grid power consumption of macro and small cells, while satisfying the MU's throughput requirement. Despite the non-convexity of the joint optimization problem, we exploit its layered structure and propose an algorithm that can efficiently compute the optimal offloading solution.
- With the optimal offloading solution for each sAP-MU pair, we further study the scenario of multiple sAPs, and investigate how the sAPs select different MUs to execute the DC-enabled traffic offloading. The formulation aims at maximizing the total network-reward that accounts for the revenue of serving the MUs' traffic requirements and the cost of the total on-grid power consumption. Despite the nature of complicated nonlinear binary programming of the formulated optimization problem, we propose an efficient layered-algorithm to solve it and find the optimal MU-selection solution.
- We present extensive numerical results to validate our proposed algorithms (for both the single-sAP case and the multi-sAP case). Moreover, we present extensive results to show the performance advantage of our proposed DC-enabled traffic offloading through the EH-powered small cells in saving the on-grid power consumption and increasing the total network-reward.

The remainder of this paper is organized as follows. We review the related studies in Section II. We present the system model and problem formulation for the single-sAP case in Section III. An efficient algorithm to compute the optimal offloading solution is proposed in Section IV. In Section V, we further study the multi-sAP case. We present the numerical results in Section VI and finally conclude this work in Section VII.

II. Related Literature

In this section, we firstly review the related studies about the energy-efficient traffic offloading in HSCNs but without considering the DC-capability. We then review the related studies that exploit the DC-capability for traffic offloading.

Studies about energy-efficient traffic offloading through HSCNs without DC: Without exploiting DC, there have been many studies investigating the energy-efficient traffic offloading through HSCNs, which can be in general categorized into two main streams.

- The first stream of studies focus on investigating the optimal resource allocations for energy-efficient traffic offloading (but without exploiting the EH power-supply) [12]–[19]. For instance, efficient schemes to optimize the heterogenous small cells' on/off mode have been proposed in [12] and [13]. Efficient schemes that optimize the tradeoff between the spectrum-efficiency and energy-efficiency have been proposed in [14] and [15]. Yu *et. al.* proposed a multi-objective optimization framework that accounts for the energy-efficiency in traffic offloading [16]. Taking into account the limited capacity of backhaul links, Yang *et. al.* proposed a refunding scheme for the small cells to accommodate the MUs offloaded from macro cells [17]. An architecture of vertical offloading has been proposed in [18] to achieve the goal of energy-saving by actively turning off the redundant cells. In [19], an energy-efficient traffic offloading scheme that exploits device-to-device communications has been proposed.
- The second stream of studies exploit the EH power-supply for traffic offloading [21]–[25]. In [21], by exploiting the statistics information about the traffic intensity and EH power-supply, Zhang *et. al.* proposed a scheme that jointly offloads the MUs to the EH-powered small cells and adjusts the small cells' on/off mode. In [22], to utilize the harvested energy, Han *et. al.* proposed a cell-size adaption scheme that actively offloads the MUs to the cells which are powered by green energy. However, aggressively offloading traffic through the EH-powered cells might lead to a severe congestion. The authors of [23] proposed a joint energy-aware and latency-aware scheme that offloads the MUs to the green-powered yet less congested small cells. In [24], by exploiting the advanced microgrids, Chia *et. al.* proposed the data offloading through the microgrid-connected small cells which are powered by EH. In [25], Chang *et. al.* proposed a wireless power transfer scheme for enabling the data offloading. In addition to the above studies targeted for traffic offloading, there are many studies focusing on the performance analysis for the EH-powered small cells [26]–[28]. In [26], Gong *et. al.* proposed a joint optimization of the cells' on-off states, resource blocks allocation, and renewable energy allocation to minimize the average on-grid power consumption. In [27], Yu *et. al.* developed a stochastic model to analyze the throughput and coverage performance when the small cells are powered by EH. In [28], Zheng *et. al.* investigated the optimal placement of the EH-assisted relay for offloading the cell-edge users' traffic. Taking into account the intermittency of EH supply as well as the environmental conditions, there have been many studies investigating the transmission outage minimization [29]–[31]. For instance, in [29], Zhou *et. al.* studied the online power control policies for outage minimization in a fading wireless link with EH-powered transmitter and receiver. In [30], Li *et. al.* proposed an optimal transmission policy for minimizing the long-term transmission outage probability, when the source node is solar-powered and equipped with

a finite-sized battery. In [31], based on the offline performance analysis, Isikman *et. al.* proposed a low-complexity online transmission scheme for optimizing the outage probability in an EH block-fading communication system.

Studies about the DC-enabled traffic offloading: Due to the benefits of enabling a flexible traffic scheduling between macro and small cells, the DC-enabled traffic offloading has attracted lots of research interests [1]–[7][1]. The functionality of DC and its performance gain have been illustrated in [1]. Exploiting the simultaneous communications with macro and small cells provided by the DC, the authors of [2] illustrated the importance of proper resource splitting in the DC-enabled transmission. Power-capacity splitting scheme for the DC-enabled traffic offloading was proposed in [3], and the traffic splitting was investigated in [4]. In [5], a joint resource allocation scheme for the DC-enabled traffic offloading has been proposed to minimize the overall resource consumption cost while satisfying each user's traffic requirement. Grouping different macro and small cells to execute the DC-enabled traffic offloading has been studied in [6]. The impact of backhaul delay in the DC-transmission has been studied in [7].

However, to the best of the authors' knowledge, few studies have investigated the DC-enabled traffic offloading through small cells which are powered by EH. As we have explained before, proper traffic scheduling and power allocation are necessitated for the DC-enabled traffic offloading through the EH-powered small cells, such that we can jointly reap the benefits of the DC and the EH power-supply.

III. System Model and Problem Formulation for One Small Cell

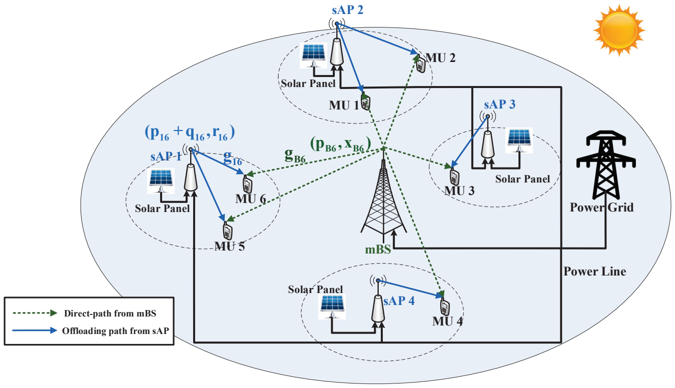


Fig. 1: An illustration of the considered DC-enabled traffic offloading model consisting of three sAPs which are powered by the hybrid energy sources and one mBS which is powered by the on-grid power-supply.

We firstly present the overall system model considered in this work, as shown in Figure 1. Specifically, a group of sAPs $\mathcal{S} = \{1, 2, ..., s, ..., S\}$ are underlaid to the coverage of a macro base station (mBS). The mBS is solely powered by the on-grid power-supply, and each sAP is powered by both the conventional on-grid power-supply and the EH power-supply. The sAPs and the mBS provide the DC-enabled downlink traffic offloading for a group of MUs $\mathcal{I} = \{1, 2, ..., i, ..., I\}$. In the following Sections III to IV, we first study the case of one sAP, by specifically focusing on investigating that one sAP (together with the mBS) provides the DC-enabled traffic offloading for one MU (i.e., one sAP-MU pair). Based on the optimal offloading solution for each individual sAP-MU pair, we further study the scenario of multiple sAPs in Section V.

A. Problem Formulation for the Pair of one sAP and one MU

We start by studying the DC-enabled traffic offloading of one sAP. To present a detailed problem formulation, we focus on investigating that the sAP (together with the mBS) provides the DC-enabled traffic offloading for one MU (i.e., one sAP-MU pair), assuming that the sAP is simultaneously offloading traffic to a total number N_s MUs. Notice that we firstly consider that the value of N_s is predetermined, and the value of N_s will be further optimized for the scenario of multiple sAPs in Section V.

For the sake of clear presentation, in the following, we denote the considered sAP as sAP s, and the MU as MU i. MU i has a traffic requirement to achieve, which is denoted by R_i^{req}. The DC enables the mBS to provide a fraction of MU i's traffic requirement and sAP s to offload the remaining part. We consider that the mBS and sAP i use different frequency channels to serve the MUs, and thus there is no co-channel interference among the mBS's transmission and the sAP's transmission. We use $p_{Bi(s)}$ to denote the mBS's transmit-power to MU i, when sAP s executes the DC-enabled offloading (notice that we include the subscript s in variable $p_{Bi(s)}$, since the mBS's transmit-power to MU i depends on which sAP executes the DC-enabled offloading). The downlink transmission-rate $x_{Bi(s)}$ from the mBS to MU i can be written as:

$$x_{Bi(s)} = W_B \log_2 \left(1 + \frac{p_{Bi(s)} g_{Bi}}{n_{Bi}}\right), \tag{1}$$

where parameter W_B denotes the mBS's downlink channel bandwidth to accommodate each MU, and g_{Bi} denotes the channel power gain from the mBS to MU i. Parameter n_{Bi} denotes the power of the background noise at MU i from the mBS's transmission, e.g., n_{Bi} can be expressed as $n_{Bi} = W_B n_0$, where n_0 is the power density of the background noise.

sAP s has both the on-grid power-supply and the EH power-supply. We use p_{si} to denote sAP s's transmit-power to MU i from its on-grid power-supply, and sAP s can flexibly adjust $p_{si} \in [0, p_s^{\max}]$ (where $p_s^{\max}$ denotes the maximum on-grid transmit-power). However, due to the intermittency of renewable energy sources, sAP s's harvested energy, which is denote by Q_s, is a random variable. As mentioned before, given the total number of N_s MUs served by sAP s, from the fairness perspective and tractability perspective, we consider that sAP s equally divides Q_s for

[1] Despite enabling the flexible traffic scheduling, the DC necessitates the coordination between the macro and small cells, and thus consumes additional resources for the signaling exchange between the macro and small cells, which however is out of the scope of our work.

all its served MUs. As a result, the offloading rate from sAP s to MU i can be written as:

$$x_{si} = W_s \log_2 \left(1 + \frac{(N_s p_{si} + Q_s) g_{si}}{N_s n_{si}} \right), \tag{2}$$

where W_s denotes the downlink channel bandwidth used by sAP s to serve each MU, and g_{si} denotes the channel power gain from sAP s to MU i. Parameter n_{si} denotes the power of the background noise at MU i from the transmission of sAP s. In this work, to focus on our objective in investigating how to properly exploit the DC-capability to facilitate the MUs' traffic offloading and mitigate the impact of the intermittency in the EH power-supply, we adopt a relatively simple assumption on the utilization of EH power-supply (i.e., sAP s equally allocates its entire EH power-supply Q_s to the served N_s MUs). Therefore, our result in this work provides a benchmark evaluation on exploiting the DC for traffic offloading when integrating the EH power-supply. As we present in this paper, such a design (i.e., our proposed joint traffic scheduling and power allocation) is already very challenging to solve, even under the current assumption. In particular, as an important direction for our future study, we will further consider that the sAP can flexibly schedule its EH power-supply over different time slots and allocate different amounts of the EH power-supply for different MUs, and investigate the corresponding optimal design of the DC-enabled traffic offloading.

In this work, we consider the long-term average on-grid power consumption and model Q_s in each short scheduling period as an independent and uniform distribution within $[M_s^{\text{low}}, M_s^{\text{upp}}]$. The values of M_s^{low} and M_s^{upp} are assumed to be known based on the historical data. The uniform distribution has been used to model uncertainty in EH power-supply over short-term periods [27], [32]–[34]. For instance, in [32], based on a time-slotted structure, the authors adopted the similar uniform distribution (as well as other distributions) to model the harvested energy from solar energy. [33] used the same slotted-structure and the uniform distribution of the EH supply to investigate the network throughput maximization for sink-based wireless sensor networks. Nevertheless, it is worth noticing that the current assumption about the EH-supply is relatively ideal for some specific cases. As an important direction for our future work, we will also consider other more realistic assumptions on the EH-supply and investigate the corresponding optimal design of DC-enabled traffic offloading.

The uncertainty in Q_s introduces randomness to the achievable offloading rate x_{si} from sAP s to MU i. Let r_{si} denote sAP s's assigned offloading rate to MU i. Due to the randomness in x_{si}, r_{si} may not be satisfied, which leads to the offloading outage. To capture this outage, we introduce function $P_{\text{out}}(p_{si}, r_{si})$ to denote the probability that sAP s's achievable offloading rate x_{si} to MU i fails to meet the assigned offloading rate r_{si}. Function $P_{\text{out}}(p_{si}, r_{si})$ can be written as:

$$\begin{aligned} P_{\text{out}}(p_{si}, r_{si}) \quad & = \mathbf{Pr}\left\{ r_{si} \geq x_{si} \right\} \\ & = \mathbf{Pr}\left\{ r_{si} \geq W_s \log_2 \left(1 + \frac{(N_s p_{si} + Q_s) g_{si}}{N_s n_{si}}\right) \right\} \\ & = \begin{cases} \frac{N_s \left((2^{\frac{r_{si}}{W_s}} - 1) \frac{n_{si}}{g_{si}} - p_{si} \right) - M_s^{\text{low}}}{M_s^{\text{upp}} - M_s^{\text{low}}}, \\ \quad \text{if } (2^{\frac{r_{si}}{W_s}} - 1) \frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s} \leq p_{si} \leq \\ \quad (2^{\frac{r_{si}}{W_s}} - 1) \frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s} \\ 0, \ \text{if } (2^{\frac{r_{si}}{W_s}} - 1) \frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s} < p_{si} \\ 1, \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

Based on $P_{\text{out}}(p_{si}, r_{si})$, we formulate an optimization problem to minimize the total on-grid power consumption, i.e., $p_{Bi(s)} + p_{si}$ when sAP s offloads traffic to MU i. The details are shown in the following total On-Grid Power Consumption Minimization (OGPM) problem:

$$\begin{aligned} & \underline{\text{(OGPM)}} && \min \ \ p_{Bi(s)} + p_{si} \\ & \text{Subject to:} && x_{Bi(s)} + r_{si}\big(1 - P_{\text{out}}(p_{si}, r_{si})\big) = R_i^{\text{req}}, && (4) \\ & && 0 \leq p_{Bi(s)} \leq p_B^{\max}, && (5) \\ & && 0 \leq p_{si} \leq p_s^{\max}, && (6) \\ & \text{Variables:} && (r_{si}, p_{si}) \ \text{and} \ \big(x_{Bi(s)}, p_{Bi(s)}\big) . \end{aligned}$$

In Problem (OGPM), we jointly optimize the following variables: i) sAP s's assigned offloading rate r_{si} and the transmit-power p_{si} to MU i, and ii) the mBS's transmission rate $x_{Bi(s)}$ and the transmit-power $p_{Bi(s)}$. Constraint (4) guarantees that MU i receives a total successful throughput equal to its requirement R_i^{req}, where the term of $r_{si}\big(1 - P_{\text{out}}(p_{si}, r_{si})\big)$ denotes the successful throughput received from sAP s. Parameter $p_B^{\max}$ denotes the mBS's maximum on-grid transmit-power for each MU, and $p_s^{\max}$ denotes sAP s's maximum on-grid transmit-power for each MU.

According to [35], Problem (OGPM) is a non-convex optimization problem which is difficult to solve. To address this difficulty, we exploit the layered-property of Problem (OGPM), and equivalently transform it into an equivalent form that can lead to Problem (OGPM) an efficient solution.

B. Layered Structure of Problem (OGPM)

We now consider the most general case of $P_{\text{out}}(p_{si}, r_{si}) \in [0, 1]$ (i.e., the first case in (3))[2]. In this case, we can

[2] We will analyze the case of full offloading outage at the end of this section and the case of zero-offloading outage in Section IV-D.

equivalently transform Problem (OGPM) into:

$$\underline{\text{(OGPM)}} \quad \min \quad p_{Bi(s)} + p_{si}$$

$$\text{Subject to: } W_B \log_2\big(1 + \frac{p_{Bi(s)} g_{Bi}}{n_{Bi}}\big) + r_{si}\frac{M_s^{\text{upp}} + N_s p_{si} - N_s(2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}}}{M_s^{\text{upp}} - M_s^{\text{low}}} = R_i^{\text{req}}, \quad (7)$$

$$(2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s} \leq p_{si} \leq (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s}, \quad (8)$$

$$\text{and constraints } (5) \text{ and } (6),$$

$$\text{Variables: } (r_{si}, p_{si}) \text{ and } p_{Bi(s)}.$$

However, constraints (7) and (8) still yield a non-convex feasible region, leading to that Problem (OGPM) is a non-convex optimization problem. To solve Problem (OGPM) efficiently, we exploit its layered-property. Specifically, we introduce an auxiliary variable $\rho_{si} \in [0, 1]$ as follows:

$$\rho_{si} = r_{si}\frac{M_s^{\text{upp}} + N_s p_{si} - N_s(2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}}}{M_s^{\text{upp}} - M_s^{\text{low}}}\frac{1}{R_i^{\text{req}}}. \quad (9)$$

Variable ρ_{si} denotes the portion of MU i's traffic requirement successfully offloaded through sAP s, and it is introduced to help us decompose Problem (OGPM) as shown in Figure 2. Specifically, we use ρ_{si} to decompose Problem (OGPM) into a top-problem (i.e., Problem (OGPM-Top)) and two parallel subproblems (i.e., Problem (Sub-mBS) and Problem (Sub-sAP)). We next explain the details about the decomposition as follows.

Given a fixed ρ_{si}, we can equivalently separate Problem (OGPM) into two parallel subproblems to minimize the BS's transmit-power and the sAP's transmit-power, respectively.

- *Subproblem to find the BS's minimum transmit-power as a function ρ_{si}:* Given ρ_{si}, the first subproblem aims at finding the BS's minimum transmit-power $\hat{p}_{Bi(s)}(\rho_{si})$ as follows:

$$\underline{\text{(Sub-mBS)}} \quad \hat{p}_{Bi(s)}(\rho_{si}) = \arg\min \quad p_{Bi(s)}$$

$$\text{Subject to: } W_B \log_2\big(1 + \frac{p_{Bi(s)} g_{Bi}}{n_{Bi}}\big) = (1 - \rho_{si})R_i^{\text{req}}, \quad (10)$$

$$\text{and constraint } (5),$$

$$\text{Variable: } p_{Bi(s)}.$$

  Notice that we denote the optimal solution $\hat{p}_{Bi(s)}(\rho_{si})$ (namely, the BS's minimum transmit-power) as a function of ρ_{si}.

- *Subproblem to find the sAP's minimum transmit-power as a function ρ_{si}* Given ρ_{si}, the second subproblem aims at finding the sAP s's minimum transmit-power $\hat{p}_{si}(\rho_{si})$ as follows:

$$\underline{\text{(Sub-sAP)}} \quad \big(\hat{p}_{si}(\rho_{si}), \hat{r}_{si}(\rho_{si})\big) = \arg\min \; p_{si}$$

$$\text{Subject to: } p_{si} = \frac{(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si} R_i^{\text{req}}}{N_s r_{si}} + (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}, \quad (11)$$

$$\text{and constraints } (6) \text{ and } (8),$$

$$\text{Variables: } p_{si} \text{ and } r_{si}.$$

  Constraint (11) stems from constraint (9). We denote the optimal solution of this subproblem (i.e., the tuple of sAP s' minimum transmit-power and assigned-offloading rate $\big(\hat{p}_{si}(\rho_{si}), \hat{r}_{si}(\rho_{si})\big)$) as a function of ρ_{si}.

By using the optimal solutions of the two subproblems at the bottom, we further optimize $\rho_{si} \in [0, 1]$ to minimize the total on-grid power consumption, which leads to Problem (OGPM-Top) as follows:

$$\underline{\text{(OGPM-Top)}} \qquad \rho_{si}^* = \arg\min \hat{p}_{Bi(s)}(\rho_{si}) + \hat{p}_{si}(\rho_{si})$$

$$\text{Variable:} \qquad 0 \leq \rho_{si} \leq 1.$$

Notice that after we solve Problem (OGPM-Top) and find ρ_{si}^*. We can express the optimal solution of Problem (OGPM) by feeding ρ_{si}^* into the two subproblems, i.e.,

$$\big(p_{Bi(s)}^*, p_{si}^*, r_{si}^*\big) = \big(\hat{p}_{Bi(s)}(\rho_{si}^*), \hat{p}_{si}(\rho_{si}^*), \hat{r}_{si}(\rho_{si}^*)\big). \quad (12)$$

In addition, $x_{Bi(s)}^* = W_B \log(1 + \frac{p_{Bi(s)}^* g_{Bi}}{n_{Bi}})$.

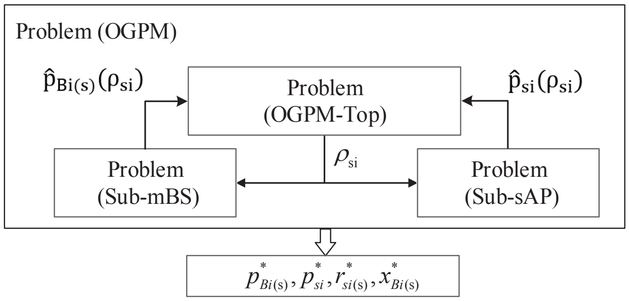

Problem (OGPM)
$\hat{p}_{Bi(s)}(\rho_{si})$
Problem (OGPM-Top)
$\hat{p}_{si}(\rho_{si})$
ρ_{si}
Problem (Sub-mBS)
Problem (Sub-sAP)
$p_{Bi(s)}^*, p_{si}^*, r_{si(s)}^*, x_{Bi(s)}^*$

Fig. 2: Decomposition of Problem (OGPM) into a top-problem (i.e., Problem (OGPM-Top)) and two subproblems (i.e., Problem (Sub-mBS) and Problem (Sub-sAP)).

We will present the detailed algorithms to solve the above three problems in the next section. Before presenting the details, we discuss a trivial case of the full offloading outage, i.e., $P_{\text{out}}(p_{si}, r_{si}) = 1$ in (3). In this case, only the mBS can provide transmission rate to MU i, and no successful offloading rate is provided by the sAP. As a result, we can directly derive the solution of Problem (OGPM) as: $x_{si,\text{F}}^* = R_i^{\text{req}}$, and $p_{Bi(s),\text{F}}^* = (2^{\frac{R_i^{\text{req}}}{W_B}} - 1)\frac{n_{Bi}}{g_{Bi}}$ (supposing that $p_{Bi(s),\text{F}}^* \leq p_B^{\max}$), and $p_{si,\text{F}}^* = 0$. Here, the subscript "F" denotes "Full Offloading Outage". Due to the triviality, we assume that the full offloading outage will not happen at the optimum of Problem (OGPM).

IV. Proposed Algorithms to Solve Problem (OGPM)

In this section, we propose algorithms to solve Problem (Sub-mBS), Problem (Sub-sAP), and Problem (OGPM-Top) in the above decomposition structure shown in Figure 2.

A. Analytical Solution of Problem (Sub-mBS)

We first solve Problem (Sub-mBS). By taking into account $p_B^{\max}$, the viable interval of ρ_{si}, which can ensure that Problem (Sub-mBS) is feasible, is $\rho_{si} \in \big[\max\{0, \rho_{si,\text{mBS}}^{\text{low}}\}, 1\big]$ where $\rho_{si,\text{mBS}}^{\text{low}} = 1 - \frac{W_B}{R_i^{\text{req}}}\log_2\Big(1 + \frac{p_B^{\max} g_{Bi}}{n_{Bi}}\Big)$. For each ρ_{si}

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2850843, IEEE Transactions on Green Communications and Networking

in this viable interval, we can derive the optimal solution of Problem (Sub-mBS) as:

$$\hat{p}_{Bi(s)}(\rho_{si}) = (2^{\frac{R_i^{\text{req}}(1-\rho_{si})}{W_B}} - 1)\frac{n_{Bi}}{g_{Bi}}, \quad (13)$$

$$\hat{x}_{Bi(s)}(\rho_{si}) = (1-\rho_{si})R_i^{\text{req}}. \quad (14)$$

B. Proposed Algorithm to solve Problem (Sub-sAP)

Problem (Sub-sAP) is difficult to solve due to the non-convexity of constraint (10). To overcome this difficulty, we transform Problem (Sub-sAP) into the following form with a single decision-variable r_{si}:

$$\underline{\text{(Sub-sAP-E)}} \quad \hat{p}_{si}(\rho_{si}) = \min \frac{(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si}R_i^{\text{req}}}{N_s r_{si}} + (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}$$

$$\text{Variable:} \quad r_{si} \geq \rho_{si}R_i^{\text{req}}. \quad (15)$$

Notice that the first "$\leq$" in constraint (8) can be directly satisfied by using (11) to replace p_{si}, and the second "$\leq$" in (8) translates to (15). Compared with Problem (Sub-sAP), we temporarily do not consider constraint (6) in Problem (Sub-sAP-E). Thanks to this temporal relaxation, Problem (Sub-sAP-E) becomes a convex optimization problem (which will be explained in Proposition 1 below). Before that, we discuss the connections between Problems (Sub-sAP) and (Sub-sAP-E).

Remark 1: Connections between Problem (Sub-sAP) and Problem (Sub-sAP-E): Problem (Sub-sAP-E) is equivalent to Problem (Sub-sAP), except that we do not include constraint (6). As a result, there will be three possible outcomes after we solve Problem (Sub-sAP-E).

- First, if the optimal solution of Problem (Sub-sAP-E) (i.e., $\hat{p}_{si}(\rho_{si})$) satisfies $0 \leq \hat{p}_{si}(\rho_{si}) \leq p_s^{\max}$, then $\hat{p}_{si}(\rho_{si})$ suffices to be the optimal solution of Problem (Sub-sAP).
- Second, if the optimal solution of Problem (Sub-sAP-E) (i.e., $\hat{p}_{si}(\rho_{si})$) satisfies $\hat{p}_{si}(\rho_{si}) > p_s^{\max}$, then Problem (Sub-sAP) is infeasible under the currently given ρ_{si}.
- Third, if $\hat{p}_{si}(\rho_{si}) < 0$, then it means that sAP s's r_{si} can be completely supported by its harvested energy, and there is no need for sAP s to use a positive on-grid power. In this case, additional operations are required such that we can find $\hat{p}_{si}(\rho_{si}) = 0$ (we will specify the details later on).

To solve Problem (Sub-sAP-E), we identify the following important property.

Proposition 1: Problem (Sub-sAP-E) is a convex optimization problem.

Proof: Let $F(r_{si})$ denote the first-order derivative of the objective function of Problem (Sub-sAP-E). We then can derive:

$$F(r_{si}) = \frac{\ln 2 n_{si}}{W_s g_{si}} 2^{\frac{r_{si}}{W_s}} - \frac{M_s^{\text{upp}} - M_s^{\text{low}}}{N_s} \frac{\rho_{si}R_i^{\text{req}}}{r_{si}^2}, \quad (16)$$

which is monotonically increasing in r_{si}. Moreover, the feasible interval of Problem (Sub-sAP-E) is affine. Thus, Problem (Sub-sAP-E) is a convex optimization problem [35]. ■

Proposition 1 enables us to use the Karush-Kuhn-Tucker (KKT) conditions [35] to compute the optimal solution of Problem (Sub-sAP-E). Specifically, we use $\hat{r}_{si}(\rho_{si})$ to denote the optimal solution of Problem (Sub-sAP-E), which depends on the given ρ_{si}. To find $\hat{r}_{si}(\rho_{si})$, we propose SubSol-Algorithm (the details are shown in the next page). The key of SubSol-Algorithm is to exploit the increasing property of $F(r_{si})$ (according to the proof of Proposition 1) and use the bisection-search to find the critical value (denoted by $r_{si}^{\text{opt,temp}}$) such that $F(r_{si}^{\text{opt,temp}}) = 0$. The WHILE-Loop (from Step 7 to Step 17) shows the bisection-search method. Since r_{si} is lower bounded by $\rho_{si}R_i^{\text{req}}$ according to (15), we directly set $r_{si}^{\text{opt,temp}} = \rho_{si}R_i^{\text{req}}$ if $F_{\text{CaseI}}(\rho_{si}R_i^{\text{req}}) > 0$, i.e., Steps 4-5 in SubSol-Algorithm. Finally, SubSol-Algorithm outputs $\hat{r}_{si}(\rho_{si}) = r_{si}^{\text{opt,temp}}$.

SubSol-Algorithm: to compute $\hat{r}_{si}(\rho_{si})$ and $\hat{p}_{si}(\rho_{si})$

```
1: Input: $\rho_{si}$.
2: MU $i$ sets $\gamma$ (i.e., the tolerable computation-error used in the bisection-search) as a very small number and sets flag $= 1$.
3: MU $i$ sets $r_{si}^{\text{lower}} = \rho_{si}R_i^{\text{req}}$ and sets $r_{si}^{\text{upper}} = r^{\text{upper}}$ (where $r^{\text{upper}}$ is a very large number).
4: if $F(r_{si}^{\text{lower}}) > 0$ then
5:     MU $i$ sets $r^{\text{opt,temp}} = r_{si}^{\text{lower}}$.
6: else
7:     while flag $= 1$ do
8:         if $(r_{si}^{\text{upper}} - r_{si}^{\text{lower}}) \leq \gamma$ then
9:             MU $i$ sets $r_{si}^{\text{opt,temp}} = \frac{1}{2}(r_{si}^{\text{lower}} + r_{si}^{\text{upper}})$ and flag $= 0$.
10:        else
11:            if $F\left(\frac{1}{2}(r_{si}^{\text{lower}} + r_{si}^{\text{upper}})\right) > 0$ then
12:                MU $i$ sets $r_{si}^{\text{upper}} = \frac{1}{2}(r_{si}^{\text{lower}} + r_{si}^{\text{upper}})$.
13:            else
14:                MU $i$ sets $r_{si}^{\text{lower}} = \frac{1}{2}(r_{si}^{\text{lower}} + r_{si}^{\text{upper}})$.
15:            end if
16:        end if
17:    end while
18: end if
19: MU $i$ computes $p_{si}^{\text{opt,temp}} = \frac{(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si}R_i^{\text{req}}}{N_s r_{si}^{\text{opt,temp}}} + (2^{\frac{r_{si}^{\text{opt,temp}}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}$.
20: if $p_{si}^{\text{opt,temp}} < 0$ then
21:     MU $i$ sets $\underline{r} = r_{si}^{\text{opt,temp}}$ and $\overline{r} = r^{\text{upper}}$.
22:     while $(\overline{r} - \underline{r}) > \gamma$ do
23:         MU $i$ sets $v = \frac{2(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si}R_i^{\text{req}}}{N_s(\underline{r} + \overline{r})} + (2^{\frac{\underline{r}+\overline{r}}{2W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}$.
24:         if $v > 0$ then
25:             MU $i$ sets $\overline{r} = \frac{1}{2}(\underline{r} + \overline{r})$.
26:         else
27:             MU $i$ sets $\underline{r} = \frac{1}{2}(\underline{r} + \overline{r})$.
28:         end if
29:     end while
30:     MU $i$ sets $r_{si}^{\text{opt,temp}} = \frac{1}{2}(\underline{r} + \overline{r})$, and $p_{si}^{\text{opt,temp}} = \frac{(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si}R_i^{\text{req}}}{N_s r_{si}^{\text{opt,temp}}} + (2^{\frac{r_{si}^{\text{opt,temp}}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}$.
31: end if
32: Output: $\hat{r}_{si}(\rho_{si}) = r_{si}^{\text{opt,temp}}$ and $\hat{p}_{si}(\rho_{si}) = p_{si}^{\text{opt,temp}}$.
```

Notice that based on $\hat{r}_{si}(\rho_{si})$, SubSol-Algorithm also outputs the smallest transmit-power required by sAP s in Step 19 (i.e., the optimal objective function value of Problem (Sub-sAP-E)) as follows:

$$\hat{p}_{si}(\rho_{si}) = \frac{(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si}R_i^{\text{req}}}{N_s \hat{r}_{si}(\rho_{si})} + (2^{\frac{\hat{r}_{si}(\rho_{si})}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}. \quad (17)$$

Viability of SubSol-Algorithm to solve Problem (Sub-sAP): We illustrate the viability of SubSol-Algorithm to solve

Problem (Sub-sAP) by addressing the three cases in Remark 1.

- First, $\big(\hat{p}_{si}(\rho_{si}), \hat{r}_{si}(\rho_{si})\big)$ (i.e., the output of SubSol-Algorithm) suffices to be the optimal solution of Problem (Sub-sAP), if $\hat{p}_{si}(\rho_{si})$ satisfies $0 \leq \hat{p}_{si}(\rho_{si}) \leq p_s^{\max}$.
- Second, Problem (Sub-sAP) is infeasible (under the given ρ_{si}), if $\hat{p}_{si}(\rho_{si})$ leads to $\hat{p}_{si}(\rho_{si}) > p_s^{\max}$.
- Third, if $\hat{p}_{si}(\rho_{si}) < 0$, it means that there is no need for sAP s to spend any positive on-grid power. In this case, we need additional operations to find the proper value of $r_{si}^{\text{opt,temp}}$ that yields $p_{si}^{\text{opt,temp}} = 0$. To this end, we design the additional Steps 20-31 in SubSol-Algorithm. We exploit the property that $\frac{(M_s^{\text{upp}} - M_s^{\text{low}})\rho_{si} R_i^{\text{req}}}{N_s r_{si}} + (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{upp}}}{N_s}$ is increasing for $r_{si} \in [r_{si}^{\text{thre}}, \infty]$ (here, r_{si}^{thre} is equal to $r_{si}^{\text{opt,temp}}$ obtained in Step 9 of SubSol-Algorithm). Hence, we use the bisection-search to find the new $r_{si}^{\text{opt,temp}}$ that yields $p_{si}^{\text{opt,temp}} = 0$.

C. Proposed Algorithm to solve Problem (OGPM-Top)

After solving Problem (Sub-mBS) and Problem (Sub-sAP) and obtaining $\hat{p}_{Bi(s)}(\rho_{si})$ and $\hat{p}_{si}(\rho_{si})$, respectively, we continue to solve Problem (OGPM-Top). In spite of its simple form, it is difficult to solve Problem (OGPM-Top), since we still cannot analytically derive $\hat{p}_{Bi(s)}(\rho_{si}) + \hat{p}_{si}(\rho_{si})$. Fortunately, Problem (OGPM-Top) only involves a single-variable ρ_{si} within a fixed interval, i.e., $\rho_{si} \in [0, 1]$. Based on this property, we propose LS-Algorithm that performs a linear-search of $\rho_{si} \in [0, 1]$ (with a very small step-size) to solve Problem (OGPM-Top) and find the optimal solution $\big(r_{si}^*, p_{si}^*, x_{Bi(s)}^*, p_{Bi(s)}^*\big)$. The details of LS-Algorithm are as follows.

LS-Algorithm: output $\big(r_{si}^*, p_{si}^*, x_{Bi(s)}^*, p_{Bi(s)}^*\big)$ for Problem (OGPM)

```
1: Initialization: Set $\rho_{si} = 0$ and $\Delta$ as a sufficiently small number ($\Delta = 10^{-5}$). MU $i$ sets the current best value CBV $= \infty$ and the current best solution CBS $= \emptyset$.
2: while $\rho_{si} \leq 1$ do
3:     If Problem (sub-sAP) is feasible, MU $i$ uses SubSol-Algorithm to compute $(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si}))$. Otherwise, turn to Step 9.
4:     If Problem (sub-mBS) is feasible, MU $i$ uses (13) and (14) to compute $\hat{x}_{Bi(s)}(\rho_{si})$ and $\hat{p}_{Bi(s)}(\rho_{si})$. Otherwise, turn to Step 9.
5:     if $\big(\hat{p}_{si}(\rho_{si}) + \hat{p}_{Bi(s)}(\rho_{si})\big) <$ CBV then
6:         MU $i$ updates CBV $= \hat{p}_{si}(\rho_{si}) + \hat{p}_{Bi(s)}(\rho_{si})$.
7:         MU $i$ sets CBS $= \big(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si}), \hat{x}_{Bi(s)}(\rho_{si}), \hat{p}_{Bi(s)}(\rho_{si})\big)$ according to (12).
8:     end if
9:     Update $\rho_{si} = \rho_{si} + \Delta$.
10: end while
11: Output: $\big(r_{si}^*, p_{si}^*, x_{Bi(s)}^*, p_{Bi(s)}^*\big) =$ CBS.
```

D. Advanced Algorithm to solve Problem (OGPM-Top) based on the Case of Zero-Outage

The linear-search in LS-Algorithm requires a very small step-size Δ (e.g., $\Delta = 10^{-5}$), which consequently requires at most $\frac{1}{\Delta}$ iterations. To reduce the number of iterations, we further propose an advanced LS-Algorithm (i.e., ADLS-Algorithm) in this subsection. The key idea is to identify a range of ρ_{si} over which we can analytically characterize the optimal solution of Problem (OGPM-Top). As a result, we do not need to use the above linear-search within this range.

1) Zero-outage Case and Its Sufficient Condition to Occur: To find such an interval of ρ_{si} over which we do not need to execute the linear-search, we identify a special case of zero-outage, namely, $P_{\text{out}}(p_{si}, r_{si}) = 0$. As an important property, we provide the following proposition.

Proposition 2: Given ρ_{si}, if $F(\rho_{si} R_i^{\text{req}}) \geq 0$, then the optimal solution of Problem (Sub-sAP-E) yields the zero-outage, i.e., $P_{\text{out}}\big(\hat{p}_{si}(\rho_{si}), \hat{r}_{si}(\rho_{si})\big) = 0$.

Proof: Based on the convexity of Problem (Sub-sAP-E), if $F(\rho_{si} R_i^{\text{req}}) \geq 0$, then the optimal solution can be directly expressed as $\hat{r}_{si}(\rho_{si}) = \rho_{si} R_i^{\text{req}}$, which consequently leads to

$$\hat{p}_{si}(\rho_{si}) = (2^{\frac{\rho_{si} R_i^{\text{req}}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s}.$$

By substituting $\big(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si})\big)$ into (3), we can obtain $P_{\text{out}}\big(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si})\big) = 0$. ■

Furthermore, we identify the following important property.

Proposition 3: There exists a critical threshold ρ_{si}^{cri}, which is given by

$$\rho_{si}^{\text{cri}} = \frac{W_s}{R_i^{\text{req}} \ln 2} \mathcal{W}(\frac{M_s^{\text{upp}} - M_s^{\text{low}}}{N_s} \frac{g_{si}}{n_{si}}). \tag{18}$$

Here, function $\mathcal{W}(.)$ is the Lambert W-function [36], i.e., the inverse function of $f(x) = x \exp(x)$. Specifically, if $\rho_{si}^{\text{cri}} < 1$, then for each $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$, the corresponding optimal solution $\big(\hat{p}_{si}(\rho_{si}), \hat{r}_{si}(\rho_{si})\big)$ of Problem (Sub-sAP-E) leads to the zero-outage.

Proof: Based on (16), we can derive $F(\rho_{si} R_i^{\text{req}}) = \frac{\ln 2}{W_s} \frac{n_{si}}{g_{si}} 2^{\frac{\rho_{si} R_i^{\text{req}}}{W_s}} - \frac{M_s^{\text{upp}} - M_s^{\text{low}}}{N_s} \frac{1}{\rho_{si} R_i^{\text{req}}}$, which is increasing in ρ_{si}. Thus, there exists a unique ρ_{si}^{cri} such that $F(\rho_{si}^{\text{cri}} R_i^{\text{req}}) = 0$. By solving $F(\rho_{si}^{\text{cri}} R_i^{\text{req}}) = 0$, we can obtain ρ_{si}^{cri} in (18). ■

2) Analytical Solution in Zero-outage Case: The purpose of analyzing the zero-outage case is that we can analytically derive the optimal solution of Problem (OGPM). The details are as follows. With $P_{\text{out}}(p_{si}, r_{si}) = 0$, Problem (OGPM) can be equivalently re-written into the following form (where, the letter "Z" denotes "Zero"):

$$\begin{array}{ll} \underline{\text{(OGPM-Z)}} & \min \;\; p_{Bi(s)} + p_{si} \\ \text{Subject to:} & W_B \log_2\big(1 + \frac{p_{Bi(s)} g_{Bi}}{n_{Bi}}\big) + r_{si} = R_i^{\text{req}} \\ & p_{si} \geq (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s} \qquad (19) \\ & \text{and constraints (5) and (6)} \\ \text{Variable:} & \{r_{si}, p_{si}\} \text{ and } p_{Bi} \end{array}$$

We next analytically derive the optimal solution of Problem (OGPM-Z). To this end, we firstly identify the following two subcases regarding the right hand side of (19):

- Subcase-I which is based on the pre-assumption that $r_{si} \leq W_s \log_2(1 + \frac{M_s^{\text{low}} g_{si}}{N_s n_{si}})$. Subcase-I means that the assigned offloading rate r_{si} is no larger than the rate that can be solely supported by sAP s's minimum EH power-supply M_s^{low}, which thus ensures the zero-outage to occur. Correspondingly, p_{si} should be zero.

- Subcase-II which is based on the pre-assumption that $r_{si} \geq W_s \log_2(1+\frac{M_s^{\text{low}} g_{si}}{N_s n_{si}})$. Subcase-II means that the assigned offloading rate r_{si} is larger than the rate that can be solely supported by sAP s's minimum EH power-supply M_s^{low}. As a result, a positive p_{si} is required to ensure the zero-outage to occur.

Based on the above Subcase-I and Subcase-II, we derive the optimal solution of Problem (OGPM-Z) under the two subcases as follows.

Solution under Subcase-I: Based on the rationale of Subcase-I, the optimal solution of Problem (OGPM-Z) can be written as:

$$\begin{aligned} r^*_{si,\text{Z-SubI}} &= \min\big\{W_s \log_2(1+\frac{M_s^{\text{low}} g_{si}}{N_s n_{si}}), R_i^{\text{req}}\big\}, \\ p^*_{si,\text{Z-SubI}} &= 0, \\ x^*_{Bi(s),\text{Z-SubI}} &= R_i^{\text{req}} - r^*_{si,\text{Z-SubI}}, \\ p^*_{Bi(s),\text{Z-SubI}} &= \big(2^{x^*_{Bi(s),\text{Z-SubI}}} - 1\big)\frac{n_{Bi}}{g_{Bi}}. \end{aligned}$$

Notice that Subcase-I is valid, only if $p^*_{Bi(s),\text{Z-SubI}} \leq p_B^{\max}$.

Solution under Subcase-II: Based on the rationale of Subcase-II, we can derive the optimal solution of Problem (OGPM-Z) as follows. Since constraint (19) is strictly binding at the optimum in this case (i.e., no additional on-grid power is required), we can equivalently transform Problem (OGPM-Z-SubII) into a single-variable optimization problem as follows:

$$\begin{aligned} &\underline{\text{(OGPM-Z-SubII)}}: \\ &\min \quad (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s} + (2^{\frac{R_i^{\text{req}} - r_{si}}{W_B}} - 1)\frac{n_{Bi}}{g_{Bi}} \\ &\text{Variable:} \quad r^{\text{low}}_{si,\text{Z-SubII}} \leq r_{si} \leq r^{\text{upp}}_{si,\text{Z-SubII}}. \end{aligned}$$

In the above problem, the lower-bound $r^{\text{low}}_{si,\text{Z-SubII}}$ is given by:

$$\begin{aligned} &r^{\text{low}}_{si,\text{Z-SubII}} = \\ &\max\left\{R_i^{\text{req}} - W_B \log_2\big(1+\frac{p_B^{\max} g_{Bi}}{n_{Bi}}\big), W_s \log_2(1+\frac{M^{\text{low}} g_{si}}{N_s n_{si}})\right\}, \end{aligned}$$

which stems from (5) and $p_{si} = (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s} \geq 0$. The upper-bound $r^{\text{upp}}_{\text{Z-SubII}}$ is given by:

$$r^{\text{upp}}_{si,\text{Z-SubII}} = \min\left\{W_s \log_2\big(1+\frac{(p_s^{\max} + \frac{M^{\text{low}}}{N_s}) g_{si}}{n_{si}}\big), R_i^{\text{req}}\right\},$$

which stems from $p_{si} = (2^{\frac{r_{si}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s} \leq p_s^{\max}$.

Notice that Subcase-II is valid, only if $r^{\text{low}}_{si,\text{Z-SubII}} \leq r^{\text{upp}}_{si,\text{Z-SubII}}$. Otherwise, Subcase-II fails to hold.

In particular, we express the optimal solution of Problem (OGPM-Z-SubII) in the following proposition.

Proposition 4: If $r^{\text{low}}_{si,\text{Z-SubII}} \leq r^{\text{upp}}_{si,\text{Z-SubII}}$, the optimal solution of Problem (OGPM-Z-SubII) can be analytically written as:

$$\begin{aligned} &r^*_{si,\text{Z-SubII}} = \\ &\begin{cases} r^{\text{low}}_{si,\text{Z-SubII}}, \ \text{if } F_{\text{Z}}(r^{\text{low}}_{si,\text{Z-SubII}}) > 0 \\ r^{\text{upp}}_{si,\text{Z-SubII}}, \ \text{if } F_{\text{Z}}(r^{\text{upp}}_{si,\text{Z-SubII}}) < 0 \\ \frac{W_B W_s}{W_B + W_s}\big(\frac{R_i^{\text{req}}}{W_B} - \log_2(\frac{W_B g_{Bi} n_{si}}{W_s g_{si} n_{Bi}})\big), \ \text{otherwise.} \end{cases} \end{aligned} \tag{20}$$

Here, $F_{\text{Z}}(r_{si}) = \frac{\ln 2 n_{si}}{W_s g_{si}} 2^{\frac{r_{si}}{W_s}} - \frac{\ln 2 n_{Bi}}{W_B g_{Bi}} 2^{\frac{R_i^{\text{req}} - r_{si}}{W_B}}$ is the first-order derivative of the objective function of Problem (OGPM-Z-SubII).

Proof: It can be verified that function $F_{\text{Z}}(r_{si})$ is increasing in r_{si}. Thus, considering the affine feasible interval, Problem (OGPM-Z-SubII) is a strictly convex optimization problem. The convexity enables us to use the KKT conditions to derive the optimal solution. Specifically, by solving $F_{\text{Z}}(r_{si}) = 0$, we obtain the last case of (20). On the other hand, if $F_{\text{Z}}(r^{\text{low}}_{si,\text{Z-SubII}}) > 0$ (which means that the objective function is increasing for $r_{si} \in [r^{\text{low}}_{si,\text{Z-SubII}}, r^{\text{upp}}_{si,\text{Z-SubII}}]$), we set $r^*_{si,\text{Z-SubII}} = r^{\text{low}}_{si,\text{Z-SubII}}$, i.e., the first case of (20). Finally, if $F_{\text{Z}}(r^{\text{upp}}_{si,\text{Z-SubII}}) < 0$ (which means that the objective function is decreasing for $r_{si} \in [r^{\text{low}}_{si,\text{Z-SubII}}, r^{\text{upp}}_{si,\text{Z-SubII}}]$), we set $r^*_{si,\text{Z-SubII}} = r^{\text{upp}}_{si,\text{Z-SubII}}$, i.e., the second case of (20). ■

By using $r^*_{si,\text{Z-SubII}}$ in (20), we derive the optimal solution of Problem (OGPM-Z-SubII) as follows:

$$\begin{aligned} p^*_{si,\text{Z-SubII}} &= (2^{\frac{r^*_{si,\text{Z-SubII}}}{W_s}} - 1)\frac{n_{si}}{g_{si}} - \frac{M_s^{\text{low}}}{N_s}, \\ x^*_{Bi(s),\text{Z-SubII}} &= R_i^{\text{req}} - r^*_{si,\text{Z-SubII}}, \\ p^*_{Bi(s),\text{Z-SubII}} &= (2^{\frac{x^*_{Bi(s),\text{Z-SubII}}}{W_B}} - 1)\frac{n_{Bi}}{g_{Bi}}. \end{aligned}$$

In summary, by comparing the optimal solutions under Subcase-I and Subcase-II (if they are feasible), we can derive the optimal solution of Problem (OGPM-Z) as follows:

$$\begin{aligned} &\big(r^*_{si,\text{Z}}, p^*_{si,\text{Z}}, x^*_{si,\text{Z}}, p^*_{Bi(s),\text{Z}}\big) = \\ &\big(r^*_{si,\text{Z-Sub}\hat{\theta}}, p^*_{si,\text{Z-Sub}\hat{\theta}}, x^*_{si,\text{Z-Sub}\hat{\theta}}, p^*_{Bi(s),\text{Z-Sub}\hat{\theta}}\big), \end{aligned} \tag{21}$$

where $\hat{\theta} = \arg\min_{\theta \in \{\text{I},\text{II}\}} p^*_{si,\text{Z-Sub}\theta} + p^*_{Bi(s),\text{Z-Sub}\theta}$.

3) Proposed Advanced LS-Algorithm: Based on Propositions 2 and 3, for the interval of $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$ (within which the optimal solution of Problem (OGPM) always leads to the zero-outage), we can directly use (21) to compute the optimal solution of Problem (OGPM), instead of executing a linear-search $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$. Exploiting this important property, we further propose the following ADLS-Algorithm ("AD" means "Advanced") to solve Problem (OGPM). The details of ADLS-Algorithm are shown on the next page.

Compared with LS-Algorithm, ADLS-Algorithm uses (21) to compute the optimal solution for the interval $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$ (i.e., Steps 3-9), and thus avoids the linear-search of $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$. In particular, let Δ denote the stepsize (which is a very small number, e.g., 10^{-5}) used by the linear-search in ADLS-Algorithm. Our proposed ADLS-Algorithm requires no more than $2\frac{\rho_{si}^{\text{cri}}}{\Delta}\log_2\big(\frac{r^{\text{upper}}}{\gamma}\big)$ (recall that γ denotes the tolerable computation-error used by our SubSol-Algorithm before). In Section VI, Figure 5(b) shows the advantage of ADLS-Algorithm in reducing the number of iterations.

Until now, we have completed solving Problem (OGPM) and obtained the optimal offloading solution for the targeted pair of sAP s and MU i, when sAP s is serving the total number of N_s MUs. Notice that by using our ADLS-Algorithm, we can find the optimal offloading solution for an arbitrary sAP-MU pair, which facilitates our extended study of the multi-sAP case in the next section.

ADLS-Algorithm: The optimal solution $\left(r^*_{si}, p^*_{si}, x^*_{Bi(s)}, p^*_{Bi(s)}\right)$ of Problem (OGPM)

```
Initialization: Set $\rho_{si} = 0$ and $\Delta$ as a sufficiently small number ($\Delta = 10^{-5}$). MU $i$ sets CBV $= \infty$ and CBS $= \emptyset$
while $\rho_{si} \leq 1$ do
    if $\rho_{si} \geq \rho^{\text{cri}}_{si}$ then
        If $r^{\text{low}}_{si,\text{Z-SubII}} \leq r^{\text{upp}}_{si,\text{Z-SubII}}$, MU $i$ computes $\left(r^*_{si,\text{Z}}, p^*_{si,\text{Z}}, x^*_{si,\text{Z}}, p^*_{Bi(s),\text{Z}}\right)$ according to (21). Otherwise, break the While-Loop.
        if $(p^*_{si,\text{Z}}+, p^*_{Bi(s),\text{Z}}) <$ CBV then
            MU $i$ updates CBV $= p^*_{si,\text{Z}}+, p^*_{Bi(s),\text{Z}}$ and CBS $= \left(r^*_{si,\text{Z}}, p^*_{si,\text{Z}}, x^*_{si,\text{Z}}, p^*_{Bi(s),\text{Z}}\right)$.
        end if
        Break the whole WHILE-LOOP.
    end if
    If Problem (sub-sAP) is feasible, MU $i$ uses SubSol-Algorithm to compute $(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si}))$. Otherwise, start the next iteration.
    If Problem (sub-mBS) is feasible, MU $i$ uses (13) and (14) to compute $\hat{x}_{Bi(s)}(\rho_{si})$ and $\hat{p}_{Bi(s)}(\rho_{si})$. Otherwise, start the next iteration.
    if $\left(\hat{p}_{si}(\rho_{si}) + \hat{p}_{Bi(s)}(\rho_{si})\right) <$ CBV then
        MU $i$ updates CBV $= \hat{p}_{si}(\rho_{si}) + \hat{p}_{Bi(s)}(\rho_{si})$.
        MU $i$ sets CBS $= \left(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si}), \hat{x}_{Bi(s)}(\rho_{si}), \hat{p}_{Bi(s)}(\rho_{si})\right)$.
    end if
    Update $\rho_{si} = \rho_{si} + \Delta$.
end while
Output: $\left(r^*_{si}, p^*_{si}, x^*_{Bi(s)}, p^*_{Bi(s)}\right) =$ CBS.
```

V. Extension to the Scenario of Multiple Small Cells

A. System Model and Problem Formulation

In this section, based on the optimal offloading solution for the single sAP case in Section IV, we further extend to investigate the scenario of multiple sAPs. As shown in the system model in Figure 1, we consider a scenario of a group of sAPs $\mathcal{S} = \{1, 2, ..., S\}$ providing the DC-enabled offloading to a group of MUs $\mathcal{I} = \{1, 2, ...I\}$. Our objective is to investigate how the sAPs properly select different MUs to provide the DC-enabled offloading, with the objective of maximizing the total network-reward. To model this problem, we introduce the binary variable $z_{si} \in \{0, 1\}, \forall i \in \mathcal{I}, s \in \mathcal{S}$ to denote whether sAP s selects MU i or not. Specifically, $z_{si} = 1$ means that sAP s selects MU i to execute the DC-enabled traffic offloading, while $z_{si} = 0$ means the opposite.

Recall that in Sections III and IV, by assuming that sAP s selects exactly $N_s = \sum_{i \in \mathcal{I}} z_{si}$ MUs to serve, we have proposed ADLS-Algorithm (and LS-Algorithm) to compute the optimal traffic scheduling and power allocation for the pair of sAP s and MU i, which is denoted by $\left(r^*_{si}, p^*_{si}, x^*_{Bi(s)}, p^*_{Bi(s)}\right)$. In other words, the optimal solution $\left(r^*_{si}, p^*_{si}, x^*_{Bi(s)}, p^*_{Bi(s)}\right)$ of Problem (OGPM) depends on the detailed value of $N_s = \sum_{i \in \mathcal{I}} z_{si}$. To explicitly denote this impact due to $\sum_{i \in \mathcal{I}} z_{si}$, in the following, we re-denote the optimal solution of Problem (OGPM) about the pair of sAP s and MU i as follows:

$$\Big(r^*_{si,(\sum_{i \in \mathcal{I}} z_{si})}, p^*_{si,(\sum_{i \in \mathcal{I}} z_{si})}, \\ x^*_{Bi(s),(\sum_{i \in \mathcal{I}} z_{si})}, p^*_{Bi(s),(\sum_{i \in \mathcal{I}} z_{si})}\Big). \tag{22}$$

Based on (22), we formulate the following optimal MU-selection problem to investigate how different sAPs optimally select different MUs to provide the DC-enabled offloading:

$$\text{(MultiMUSel):} \ \max \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{I}} \Big(\mu R^{\text{req}}_i - \pi(p^*_{si,(\sum_{i \in \mathcal{I}} z_{si})} + \\ p^*_{Bi(s),(\sum_{i \in \mathcal{I}} z_{si})})\Big) z_{si}$$

$$\text{Subject to:} \ \sum_{s \in \mathcal{S}} z_{si} \leq 1, \forall i \in \mathcal{I} \tag{23}$$

$$\sum_{i \in \mathcal{I}} z_{si} \leq H^{\max}_s, \forall s \in \mathcal{S} \tag{24}$$

$$z_{si} = 0, \ \text{if} \ i \in \Omega^{\text{inf}}_{s,(\sum_{i \in \mathcal{I}} z_{si})}, \forall s \in \mathcal{S}, i \in \mathcal{I} \tag{25}$$

$$\text{Variables:} \ \{z_{si}\}_{s \in \mathcal{S}, i \in \mathcal{I}}.$$

In Problem (MultiMUSel), we aim at maximizing the total network-reward that takes into account the marginal reward λ for successfully serving a MU's traffic requirement, and the cost due to the mBS's and sAP s's total on-grid power consumption $\left(p^*_{si,(\sum_{i \in \mathcal{I}} z_{si})} + p^*_{Bi(s),(\sum_{i \in \mathcal{I}} z_{si})}\right)$ when sAP s selects MU i to provide the traffic offloading. Here, parameter π denotes the marginal cost for the on-grid power consumption. Constraint (23) means that MU i can only be served by at most one sAP. Constraint (24) means that sAP s can select no more than $H^{\max}_s$ MUs to serve. Here, we consider that the small cells use the frequency division multiple access (FDMA) to accommodate different MUs, and each sAP s has $H^{\max}_s$ available sub-channels to serve the MUs. In constraint (25), set $\Omega^{\text{inf}}_{s,(\sum_{i \in \mathcal{I}} z_{si})}$ denotes the subset of the MUs who cannot be served by sAP s when sAP s selects total $\sum_{i \in \mathcal{I}} z_{si}$ MUs to serve[3].

Problem (MultiMUSel) is very challenging to solve, since it is a nonlinear binary programming problem due to the following two reasons. First, in the objective function, for each pair of sAP s and MU i, the minimum on-grid power $\left(p^*_{si,(\sum_{i \in \mathcal{I}} z_{si})}, p^*_{Bi(s),(\sum_{i \in \mathcal{I}} z_{si})}\right)$ depends on the value of $\sum_{i \in \mathcal{I}} z_{si}$. Second, constraints (23) and (24) together lead to a resource-constrained generalized assignment problem [38]. Specifically, each sAP s (i.e., an agent) can accept no more than $H^{\max}_s$ MUs (i.e., the jobs), and each MU i can only be assigned to at most one sAP. To tackle this difficulty, we propose an efficient algorithm to solve Problem (MultiMUSel) in the next subsection.

B. Layered Algorithm to Solve Problem (MultiMUSel)

To solve Problem (MultiMUSel), we first identify the following property: in Problem (MultiMUSel), the objective function and constraints are separable with respect to individual sAP, except that constraint (23) couples all sAPs. To decouple (23), for each sAP s, we first introduce set $\Lambda_s \subseteq \mathcal{I}$ to denote the subset of the MUs who are assigned to sAP s as *the candidate-users to be served*. Please notice that the MUs in Λ_s are the candidate-users to be selected by sAP s (in other words, it might be optimal for sAP s to only select some MUs in Λ_s, instead of all of them). In addition, we introduce Λ_0 to denote the subset of MUs who are not

[3] Notice that we can determine set $\Omega^{\text{inf}}_{s,(\sum_{i \in \mathcal{I}} z_{si})}$ as follows. Given the value of $\sum_{i \in \mathcal{I}} z_{si}$, MU i belongs to set $\Omega^{\text{inf}}_{s,(\sum_{i \in \mathcal{I}} z_{si})}$, if we find that Problem (OGPM) is infeasible for the pair of sAP s and MU i.

assigned to any sAP (i.e., the MUs Λ_0 will be not served by any sAP). We impose the following two constraints regarding $\{\Lambda_s\}_{s\in\mathcal{S}\cup\{0\}}$, i.e., i) $\bigcup_{s\in\mathcal{S}\cup\{0\}}\Lambda_s=\mathcal{I}$, and ii) $\Lambda_s\bigcap\Lambda_{s'}=\emptyset$ for any two different s and s'.

Based on $\{\Lambda_s\}_{s\in\mathcal{S}\cup\{0\}}$, our key idea to solve Problem (MultiMUSel) is to vertically decompose it into the following two problems as shown in Figure 3. We explain the details about the decomposition as follows.

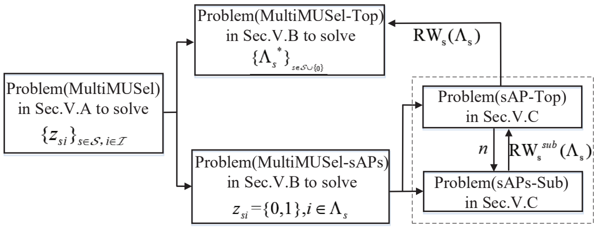

Problem(MultiMUSel)
in Sec.V.A to solve
$\{z_{si}\}_{s\in\mathcal{S},\, i\in\mathcal{I}}$
Problem(MultiMUSel-Top)
in Sec.V.B to solve
$\{\Lambda_s^*\}_{s\in\mathcal{S}\cup\{0\}}$
$\mathrm{RW_s}(\Lambda_s)$
Problem(MultiMUSel-sAPs)
in Sec.V.B to solve
$z_{si}=\{0,1\}, i\in\Lambda_s$
Problem(sAP-Top)
in Sec.V.C
n
$\mathrm{RW_s}^{sub}(\Lambda_s)$
Problem(sAPs-Sub)
in Sec.V.C

Fig. 3: Decomposition of Problem (MultiMUSel) into Problem (MultiMUSel-Top) on the top and Problem (MultiMUSel-sAPs) for each sAP s at the bottom. Problem (MultiMUSel-sAPs) is further decomposed into Problem (sAP-Top) and Problem (sAP-Sub) in Section V-C.

1) Subproblem to optimize the MU-selection for each individual sAP under given Λ_s: Suppose that $\{\Lambda_s\}_{s\in\mathcal{S}\cup\{0\}}$ is given. We solve the following subproblem for each sAP s:

(MultiMUSel-sAPs):

$$\mathrm{RW}_s(\Lambda_s)=\max\sum_{i\in\Lambda_s}\big(\mu R_i^{\mathrm{req}}-\pi(p^*_{si,(\sum_{i\in\Lambda_s}z_{si})}+p^*_{Bi(s),(\sum_{i\in\Lambda_s}z_{si})})\big)z_{si}$$

$$\text{Subject to: }\sum_{i\in\Lambda_s}z_{si}\le H_s^{\max}\tag{26}$$

$$z_{si}=0,\ \text{ if } i\in\Omega^{\mathrm{inf}}_{s,(\sum_{i\in\Lambda_s}z_{si})}\tag{27}$$

$$\text{Variables: } z_{si}=\{0,1\}, i\in\Lambda_s.$$

As we will show later on, we can solve Problem (MultiMUSel-sAPs) and derive $\mathrm{RW}_s(\Lambda_s)$ efficiently.

2) Top-problem to optimize $\{\Lambda_s\}_{s\in\mathcal{S}}$ for all sAPs: After solving Problem (MultiMUSel-sAPs) and obtaining $\mathrm{RW}_s(\Lambda_s)$ for each sAP s, we then continue to find the optimal $\{\Lambda_s^*\}_{s\in\mathcal{S}\cup\{0\}}$, by solving the following optimization problem:

$$\begin{aligned}\text{(MultiMUSel-Top):}\quad & \max\sum_{s\in\mathcal{S}}\mathrm{RW}_s(\Lambda_s)\\ \text{Subject to:}\quad & \bigcup_{s\in\mathcal{S}\cup\{0\}}\Lambda_s=\mathcal{I}\\ & \Lambda_s\bigcap\Lambda_{s'}=\emptyset,\forall s\neq s'\\ \text{Variables:}\quad & \{\Lambda_s\}_{s\in\mathcal{S}\cup\{0\}}.\end{aligned}$$

Notice that after obtaining $\{\Lambda_s^*\}_{s\in\mathcal{S}\cup\{0\}}$, we can obtain the optimal MU-selection solution for the original Problem (MultiMUSel), i.e., by solving Problem (MultiMUSel-sAPs) for each sAP s again under the given Λ_s^*.

In the following, we provide the algorithms to solve Problem (MultiMUSel-sAPs) and Problem (MultiMUSel-Top), respectively.

C. A Further Decomposition of Problem (MultiMUSel-sAPs)

Firstly, we focus on solving Problem (MultiMUSel-sAPs) under the given Λ_s. However, Problem (MultiMUSel-sAPs) is still a nonlinear binary programming problem. To efficiently solve Problem (MultiMUSel-sAPs), we further decompose it into two subproblems (as shown in Figure 3). Specifically, we introduce a variable n to denote the number of the MUs selected by sAP s to serve, and the feasible value of n is within $1,2,...,|\Lambda_s|$ (here, $|\Lambda_s|$ denotes the cardinality of set Λ_s which is given in advance in Problem (MultiMUSel-sAPs)). By using the newly introduced variable n, we present the vertical decomposition of Problem (MultiMUSel-sAPs) as follows.

1) Subproblem (sAPs-Sub) to optimize $\{z_{si}\}_{i\in\Lambda_s}$ under given n: Suppose that the value of n is given in advance, which means that sAP s selects $\sum_{i\in\Lambda_s}z_{si}=\min\{H_s^{\max},n\}$ MUs in Λ_s. In this situation, we focus on solving the following subproblem under the given n (as well as the given Λ_s):

(sAPs-Sub): $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s,n)=$

$$\max\sum_{i\in\Lambda_s}\big(\mu R_i^{\mathrm{req}}-\pi(p^*_{si,(\min\{H_s^{\max},n\})}+p^*_{Bi(s),(\min\{H_s^{\max},n\})})\big)z_{si}$$

$$\text{Subject to: }\sum_{i\in\Lambda_s}z_{si}=\min\{H_s^{\max},n\},$$

$$z_{si}=0,\ \text{ if } i\in\Omega^{\mathrm{inf}}_{s,(\min\{H_s^{\max},n\})},$$

$$\text{Variables: } z_{si}=\{0,1\}, i\in\Lambda_s.$$

Since the value of $\min\{H_s^{\max},n\}$ is known, Problem (sAPs-Sub) is a linear binary programming problem, which differs from Problem (MultiMUSel-sAPs). We will show that we can analytically derive $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s,n)$ and thus solve Problem (sAPs-Sub).

2) Top-problem (sAPs-Top) to optimize n: After obtaining $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s,n)$ for the given n, we then solve the following problem to find the optimal n^* that can maximize sAP s's reward under the given set Λ_s:

(sAPs-Top):

$$\mathrm{RW}_s(\Lambda_s)=\max_{n=\{0,1,2,...,\min\{|\Lambda_s|,H_s^{\max}\}\}}\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s,n).\tag{28}$$

Notice that after solving Problem (sAPs-Top), we complete solving the original Problem (MultiMUSel-sAPs)

D. Proposed Algorithm to Problem (MultiMUSel-sAPs) for each sAP s

Based on the decomposition of Problem (MultiMUSel-sAPs) explained in the previous subsection, we next solve Problem (MultiMUSel-sAPs). Specifically, we first analytically solve Problem (sAPs-Sub), and then propose an algorithm to solve Problem (sAPs-Top) by using the analytical solution of Problem (sAPs-Sub).

1) Analytical solution of Problem (sAPs-Sub): We first focus on solving Problem (sAPs-Sub) and deriving $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s,n)$. Thanks to the simple structure of Problem (sAPs-Sub), we can derive the optimal solution as follows.

Specifically, suppose that the MUs in Λ_s are ordered in the descending order based on the value of $V_m = \mu R_m^{\mathrm{req}} - \pi(p^*_{sm,(\min\{H_s^{\max},n\})} + p^*_{Bm(s),(\min\{H_s^{\max},n\})})$, i.e.

$$V_1 > V_2 > V_3 > ... > V_{|\Lambda_s|}. \tag{29}$$

In the reminder of this section, we assume that MU m in Λ_s has been ordered according to (29) when we use subscript m to denote the MU. We provide the following result regarding the optimal solution of Problem (sAPs-Sub).

Proposition 5: Based on the ordering in (29), the optimal solution of Problem (sAPs-Sub) (with the given $n \leq \min\{|\Lambda_s|, H_s^{\max}\}$) can be given by

$$z^*_{sm,(n)} = \begin{cases} 1, \text{ if } m \notin \Omega^{\mathrm{inf}}_{s,(\min\{H_s^{\max},n\})} \text{ and} \\ \quad \sum_{j\in\Lambda_s, j=1}^{j\in\Lambda_s, j=m-1} z^*_{sj} < \min\{H_s^{\max}, n\} \\ 0, \text{ otherwise} \end{cases}$$

Please notice that the subscript (n) in $z^*_{sm,(n)}$ indicates that the optimal solution depends on the given value of n. On the other hand, Problem (sAPs-Sub) is infeasible, if the following condition holds:

$$|\Lambda_s| - |\Omega^{\mathrm{inf}}_{s,(\min\{H_s^{\max},n\})}| < \min\{H_s^{\max}, n\}. \tag{30}$$

Proof: Due to the structure of Problem (sAPs-Sub), we can prove (30) by showing the contradiction. Let $\{z^*_{sm}\}_{m\in\Lambda_s}$ denote the optimal solution of Problem (sAPs-Sub) but being inconsistent with (30). In other words, there exists two different m and m' (with m and $m' \in \Lambda_s$, $m' > m$, m and $m' \notin \Omega^{\mathrm{inf}}_{s,(\min\{H_s^{\max},n\})}$), and we have $z^*_{sm} = 0$ and $z^*_{sm'} = 1$, which is inconsistent with (30). In this situation, we can set $z^*_{sm} = 1$ and $z^*_{sm'} = 0$ to increase the objective function of Problem (sAPs-Sub) but without violating any constraint. We thus finish the proof. ■

Based on Proposition 5 and (30), we can express the optimal objective function value of Problem (sAPs-Sub) as follows

$$\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s, n) = \sum_{m\in\Lambda_s} \big(\mu R_i^{\mathrm{req}} - \pi(p^*_{sm,(\min\{H_s^{\max},n\})} + p^*_{Bm(s),(\min\{H_s^{\max},n\})})\big) z^*_{sm,(n)}. \tag{31}$$

2) Solving Problem (sAPs-Top) and Problem (MultiMUSel-sAPs): Based on (30) and (31), we then solve Problem (sAPs-Top). Since Problem (sAPs-Top) only involves an integer variable $n = \{0, 1, 2, ..., \min\{|\Lambda_s|, H_s^{\max}\}\}$, we propose sAPSol-Algorithm, which is based on the enumeration of n within $\{0, 1, 2, ..., \min\{|\Lambda_s|, H_s^{\max}\}\}$, to find the optimal n^* that can maximize the objective function $\mathrm{RW}_s(\Lambda_s)$. Notice that our proposed sAPSol-Algorithm also solves Problem (MultiMUSel-sAPs) and outputs the optimal solution $\{z^*_{sm}\}_{m\in\Lambda_s}$ (as well as the corresponding $\mathrm{RW}_s(\Lambda_s)$).

E. Proposed Algorithm to solve Problem (MultiMUSel-Top)

By using sAPSol-Algorithm as the subroutine to compute $\mathrm{RW}_s(\Lambda_s)$ (for each sAP s) under the given Λ_s, we then continue to solve Problem (MultiMUSel-Top). Thanks to the simple form, Problem (MultiMUSel-Top) can be considered as an optimal grouping problem that assigns the MUs into the sets $\{\Lambda_s\}_{s\in\mathcal{S}\bigcup\{0\}}$. We use $\{\Lambda^*_s\}_{s\in\mathcal{S}\bigcup\{0\}}$ to denote the optimal solution of Problem (MultiMUSel-Top). To find $\{\Lambda^*_s\}_{s\in\mathcal{S}\bigcup\{0\}}$, we propose the following SelSol-Algorithm. The key of SelSol-Algorithm is to execute a randomized local search based on the idea of simulated annealing [39].

sAPSol-Algorithm: to solve Problem (MultiMUSel-sAPs) and output $\{z^*_{sm}\}_{m\in\Lambda_s}$ and $\mathrm{RW}_s(\Lambda_s)$

```
Input $\Lambda_s$ for sAP $s$
Initialize CBV = 0, CBS = $\emptyset$, and $n = 0$.
while $n \leq \min\{|\Lambda_s|, H_s^{\max}\}$ do
    sAP $s$ uses ADLS-Algorithm to compute $(p^*_{si,(n)}, p^*_{Bi(s),(n)})$ for each MU $i \in \Lambda_s$, and obtain $\Omega^{\mathrm{inf}}_{s,(n)}$.
    if Problem (sAP$s$-Sub) is infeasible according to (30) then
        sAP sets $n = n + 1$ and starts the next round of iteration.
    else
        sAP $s$ uses (30) to compute $\{z^*_{sm,(n)}\}_{m\in\Lambda_s}$ and uses (31) to compute $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s, n)$.
        if $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s, n) >$ CBV then
            sAP $s$ sets CBV = $\mathrm{RW}_s^{\mathrm{sub}}(\Lambda_s, n)$, and set CBS = $\{z^*_{sm,(n)}\}_{m\in\Lambda_s}$.
        end if
        sAP $s$ sets $n = n + 1$.
    end if
end while
Output: $\mathrm{RW}_s(\Lambda_s)$ = CBV and $\{z^*_{sm}\}_{m\in\Lambda_s}$ = CBS.
```

SelSol-Algorithm: to solve Problem (MultiMUSel-Top) and output $\{\Lambda^*_s\}_{s\in\mathcal{S}\bigcup\{0\}}$

```
Initialization: assign the MUs into $\{\Lambda\}_{s\in\mathcal{S}\bigcup\{0\}}$ in a round-robin manner, set the iteration index $t = 1$, and set the initial temperature $T_{ini} = 100$.
Each sAP $s$ uses sAPSol-algorithm to compute $\mathrm{RW}_s(\Lambda_s)$, and the virtual sAP 0 sets $\mathrm{RW}_0(\Lambda_0) = 0$ directly.
while (1) do
    Randomly select an sAP (let us say $s$) with nonempty $\Lambda_s$, and sAP $s$ randomly selects a MU $j \in \Lambda_s$. sAP $s$ further randomly selects another sAP $s' \neq s$.
    sAP $s$ moves MU $j$ to sAP $s'$. Correspondingly, sAP $s$ sets $\widetilde{\Lambda}_s = \Lambda_s \setminus \{j\}$, and sAP $s$ sets $\widetilde{\Lambda}'_s = \Lambda'_s \bigcup\{j\}$.
    if $\mathrm{RW}_s(\widetilde{\Lambda}_s) + \mathrm{RW}_s(\widetilde{\Lambda}'_s) > \mathrm{RW}_s(\Lambda_s) + \mathrm{RW}_s(\Lambda'_s)$ then
        sAP $s$ updates $\Lambda_s = \widetilde{\Lambda}_s$, and sAP $s'$ updates $\Lambda'_s = \widetilde{\Lambda}'_s$.
    else
        With probability equal to $\exp\{\frac{\Delta}{\kappa T_t}\}$ where $\Delta = \mathrm{RW}_s(\widetilde{\Lambda}_s) + \mathrm{RW}_s(\widetilde{\Lambda}'_s) - \mathrm{RW}_s(\Lambda_s) - \mathrm{RW}_s(\Lambda'_s)$, and $T_t = \frac{T_{\mathrm{ini}}}{1+\alpha * t^2}$ is the system temperature at time $t$, $\kappa$ is the Bolzmann constant. sAP $s$ updates $\Lambda_s = \widetilde{\Lambda}_s$, and sAP $s'$ updates $\Lambda'_s = \widetilde{\Lambda}'_s$.
    end if
    if the set of $\{\Lambda_s\}_{s\in\mathcal{S}}$ do not change for consecutive iterations then
        Reach convergence and break the WHILE-LOOP.
    end if
    Update $t = t + 1$.
end while
Output $\Lambda^*_s = \Lambda_s, \forall s \in \mathcal{S}$.
```

- In each round of iteration, a randomly selected sAP s randomly selects a MU $j \in \Lambda_s$ and moves this MU j to another another randomly selected sAP s'. If such a MU-switch can improve the total reward of sAP s and sAP s', then sAP s and sAP s' accept such a MU-switch by updating $\Lambda_s = \Lambda_s \backslash \{j\}$ and $\Lambda_{s'} = \Lambda_{s'} \bigcup \{j\}$. Please notice that for the sake of easy presentation, we introduce sAP 0 as a virtual sAP which manages Λ_0.
- To avoid being trapped in the local optimum, we adopt the idea of Simulated Annealing (SA) [39], [40] to accept the non-improvement MU-switch with a certain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2850843, IEEE Transactions on Green Communications and Networking

probability (in Step 4). In particular, the probability to accept the non-improvement MU-switch depends on both the reward degradation and the current temperature (i.e., T_t). Specifically, we use the cooling schedule $T_t = \frac{T_{\text{ini}}}{1+\alpha * t^2}$ (with T_{ini} denoting the initial temperature, $\alpha > 0$ being a constant, and t denoting the iteration index) [41][4]. The higher the temperature, the more likely to accept the non-improvement exchange. When the temperature decreases, the probability to accept the non-improvement MU-switch decreases.

After obtaining $\{\Lambda_s^*\}_{s \in \mathcal{S} \bigcup \{0\}}$, we can finally compute the optimal MU-selection solution for the original Problem (MultiMUSel) (in Section V-A), by executing sAPSol-Algorithm for each sAP s under the given Λ_s^*.

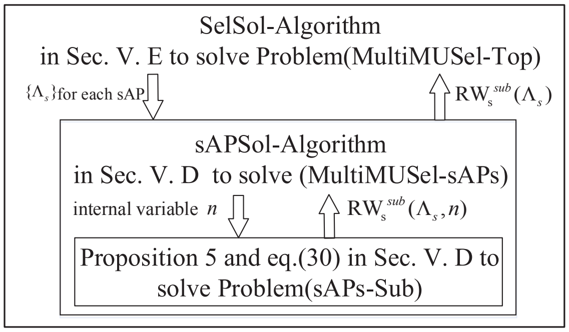

SelSol-Algorithm
in Sec. V. E to solve Problem(MultiMUSel-Top)
$\{\Lambda_s\}$for each sAP
$\text{RW}_\text{s}^{sub}(\Lambda_s)$
sAPSol-Algorithm
in Sec. V. D to solve (MultiMUSel-sAPs)
internal variable n
$\text{RW}_\text{s}^{sub}(\Lambda_s, n)$
Proposition 5 and eq.(30) in Sec. V. D to solve Problem(sAPs-Sub)

Fig. 4: Connections between SelSol-Algorithm, sAPSol-Algorithm, and Proposition 5 and eq. (31)

As a summary of in this section, we provide Figure 4 to illustrate the connections among our proposed algorithm, and more importantly, how they work together to find the optimal solution of Problem (MultiMUSel).

VI. Numerical Results

A. Numerical Results for the Single sAP Case

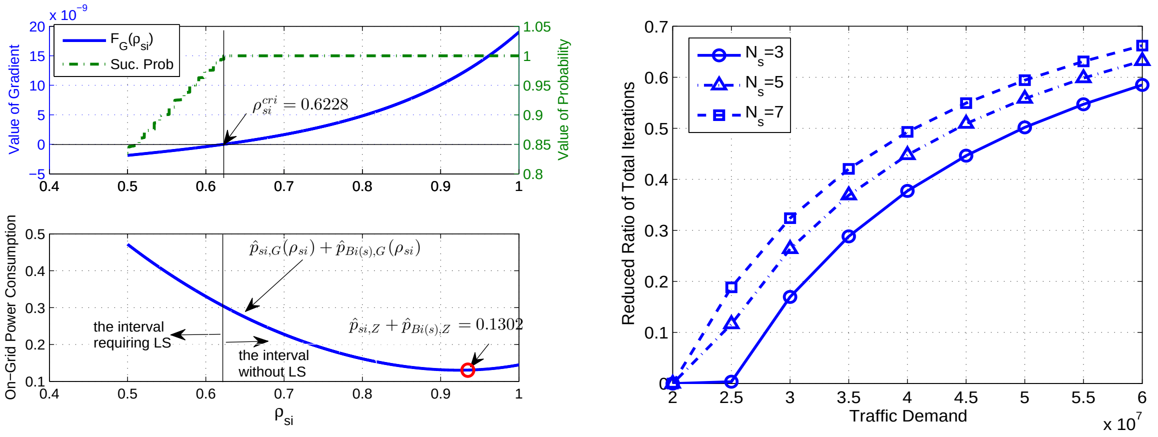
We first validate our analytical results and the proposed algorithms for the case of one sAP. We consider a scenario in which the mBS is located at the origin $(0\text{m}, 0\text{m})$, and sAP s is located at $(250\text{m}, 0\text{m})$. The representative MU i, which forms the DC-pair with sAP s is located at $(220\text{m}, 0\text{m})$ (later on we will specify the value of N_s, i.e., the number of the MUs served by sAP s). We set the channel power gain g_{Bi} from the mBS to from MU i according to the path-loss model, i.e., $g_{Bi} = \lambda d_{Bi}^{-\varphi}$, in which parameter d_{Bi} denotes the distance between the mBS and MU i, parameter φ denotes the scaling-parameter (we use $\varphi = 2.5$), and λ follows an exponential distribution with the unit mean for capturing the impact of channel fading. The channel power gain g_{si} from sAP s to MU i is generated in a similar way. With this setting, the randomly generated channel power gains are $g_{Bi} = 6.383 \times 10^{-7}$ and $g_{si} = 8.620 \times 10^{-5}$, which are used in the following Figures 5 to 7. In addition, we set $p_{Bi}^{\max} = 1\text{W}$ and $p_{si}^{\max} = 0.4\text{W}$ [37], and set the bandwidths $W_B = 10\text{MHz}$, $W_s = 5\text{MHz}$, and set $n_0 = 10^{-14}\text{W}$.

Verification of ADLS-Algorithm: Figure 5(a) shows the operations of ADLS-Algorithm that enumerates ρ_{si} for solving Problem (OGPM). We set $N_s = 3$. Recall that for each enumerated ρ_{si}, we use SubSol-Algorithm to compute $(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si}))$, and use (13) and (14) to compute $\hat{x}_{Bi(s)}(\rho_{si})$ and $\hat{p}_{Bi(s)}(\rho_{si})$, respectively. In addition, we use (18) to compute $\rho_{si}^{\text{cri}} = 0.6228$ (which is marked out in Figure 5(a)). The top-subplot of Figure 5(a) shows that $(\hat{r}_{si}(\rho_{si}), \hat{p}_{si}(\rho_{si}))$ leads to the zero-outage when $\rho_{si} \geq \rho_{si}^{\text{cri}}$, which thus validates Proposition 2. The bottom-subplot of Figure 5(a) shows the on-grid power consumption $\hat{p}_{si}(\rho_{si}) + \hat{p}_{Bi(s)}(\rho_{si})$ when enumerating ρ_{si}. Meanwhile, we use (21) to compute $p_{si,\text{Z}}^* + p_{Bi(s),\text{Z}}^*$, which is marked out by the red circle. In particular, $p_{si,\text{Z}}^* + p_{Bi(s),\text{Z}}^*$ exactly corresponds to the minimum of $\hat{p}_{si}(\rho_{si}) + \hat{p}_{Bi(s)}(\rho_{si})$ for the interval of $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$. This validates our analysis for the zero-outage case and our proposed ADLS-Algorithm, i.e., directly calculating $p_{si,\text{Z}}^* + p_{Bi(s),\text{Z}}^*$, instead of using the linear-search for $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$.

Advantage of of ADLS-Algorithm: Figure 5(b) shows the advantage of ADLS-Algorithm in reducing the iterations, in comparison with LS-Algorithm. As stated in Section IV, by using the analytical solution (21) for the zero-outage case, ADLS-Algorithm can avoid the linear-search of $\rho_{si} \in [\rho_{si}^{\text{cri}}, 1]$, which thus reduces the number of required iterations. Specifically, we plot the ratio of reduced iterations (i.e., the value of $1 - \rho_{si}^{\text{cri}}$) by using ADLS-Algorithm in Figure 5(b). Figure 5(b) shows that the reduced ratio increases quickly in both MU i's traffic requirement and the number of the MUs served by sAP s. This result means that our ADLS-Algorithm is more computationally efficient when the MU's traffic requirement (or the total number of the MUs served by the sAP) is larger.

Illustration of Optimal Offloading Solution: Figure 6(a) illustrates the optimal solution of Problem (OGPM) versus different traffic requirements. Here, we set $N_s = 7$ (i.e., the sAP is serving 7 MUs). The top-subplot of Figure 6(a) plots sAP s's optimal on-grid transmit-power, the mBS's optimal transmit-power, and the minimum total on-grid power consumption. As shown in the top-subplot of Figure 6(a), when the MU's traffic requirement is low, the minimum total on-grid power consumption is zero. This is because that we can completely rely on the sAP's EH power-supply to power the offloading in order to meet the MU's traffic requirement. However, when the MU's traffic requirement increases, the sAP's EH power-supply alone cannot satisfy the MU's requirement. Thus, the sAP needs to spend a non-zero on-grid power to afford the MU's requirement, which yields the increase in the sAP's optimal on-grid power. Moreover, when the MU's traffic requirement further increases, traffic offloading through the sAP will consume a large on-grid power. As a result, the mBS needs to spend a non-zero on-grid power to afford part of the MU's traffic requirement, which yields the increase in the mBS's optimal on-grid

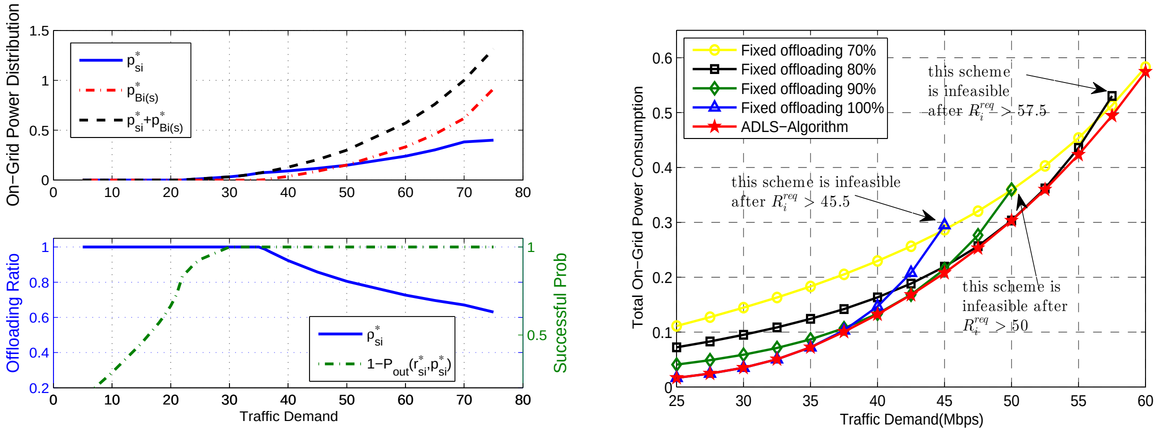
[4] This cooling schedule can yield an asymptotic convergence to the global optimum solution according to [42]. Specifically, based on Theorem 1 in [42], a cooling scheme can yield an asymptotic convergence to the global optimum solution, if the following conditions are satisfied, i.e., (i) $\lim_{t\to\infty} T(t) = 0$, and (ii) $\sum_{t=1}^{\infty} \exp[-\frac{d^*}{T(t)}] = \infty$, where d^* can be regarded as the distance between the optimal solution and other ones. In particular, in our Problem (MultiMUSel-Top), the value of d^* is a finite yet fixed number. Thus, we can show that the adopted cooling schedule $T_t = \frac{T_{\text{ini}}}{1+\alpha * t^2}$ fits the two aforementioned conditions, which can yield the asymptotic convergence to the global optimum solution.



(a) An illustrative example of executing ADLS-Algorithm. Top: Example of $F(r_{si})$. Bottom: Total on-grid power.

(b) Advantage of reducing the number of iterations used by ADLS-Algorithm in comparison with using LS-Algorithm.

Fig. 5: Performance of ADLS-Algorithm. Left: Example of executing ADLS-Algorithm. Right: Advantage of ADLS-Algorithm.



(a) Illustration of the optimal solution of Problem (OGPM).

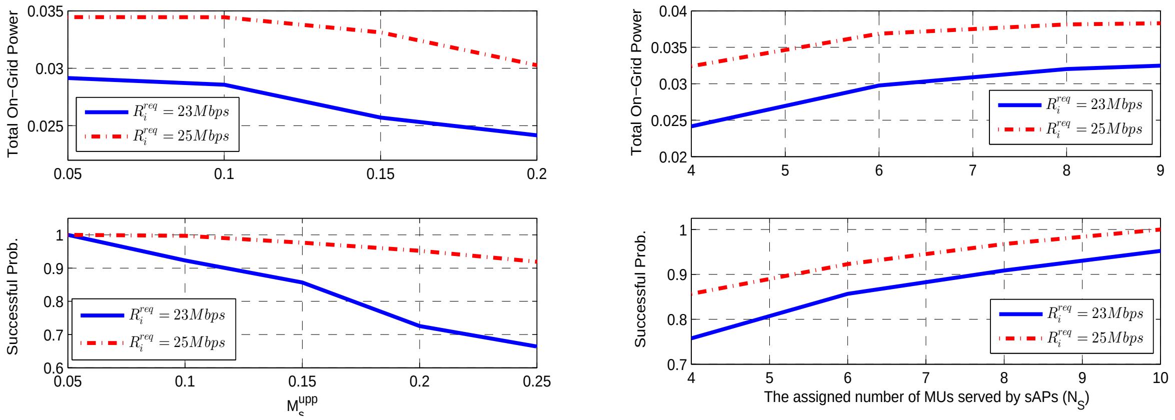
(b) Advantage in saving the total on-grid power consumption.

Fig. 6: Illustration of the optimal offloading solution and the advantage in reducing the total on-grid power consumption.

power consumption. In this situation, the corresponding optimal offloading-ratio (i.e., the value of $1 - x^*_{Bi(s)}/R^{\text{req}}$) starts to decrease, as shown in the bottom-subplot of Figure 6(a). In particular, the bottom-subplot of Figure 6(a) also shows that the successful offloading probability (i.e., the value of $1 - P_{\text{out}}(r^*_{si}, p^*_{si})$) gradually increases to one, when the MU's traffic requirement increases. The reason is that the throughput offloaded through the sAP increases, when the MU's traffic requirement increases. As a result, the sAP needs to be more conservative in relying on the EH power-supply, but using more on-grid power to support the traffic offloading. This is essentially because that a larger offloading outage probability will lead to a larger waste of the sAP's on-grid power consumption.

Advantage of Optimal Offloading Solution: Figure 6(b) further shows the advantage of the proposed optimal offloading scheme in reducing the total on-grid power consumption. For the purpose of comparison, we also consider another offloading scheme in which the MU offloads a fixed portion of its traffic requirement to the sAP (we set such a portion as 70%, 80%, 90%, and 100% in Figure 6(b)). The results validate the advantage of our proposed offloading scheme, i.e., it can minimize the total on-grid power consumption while guaranteeing the served MU's traffic requirement. This advantage is essentially achieved by our formulated joint optimization of the traffic scheduling and power allocation, which is able to jointly reap the benefit of DC-capability (to flexibly schedule the MU's traffic between macro and small cells) and the benefit of exploiting EH power-supply (to reduce the on-grid power consumption). In comparison, the fixed offloading scheme fails to achieve these benefits. Specifically, as shown in Figure 6(b), due to the fact the sAP's EH power-supply cannot accommodate a very large offloading rate, offloading too much of the MU's traffic (i.e., the 100%-offloading) will lead to a quick increase in the total on-grid power consumption when the MU's traffic requirement increases.

Impact of the EH power-supply: To evaluate the impact of the EH power-supply, we plot the optimal solution versus different degrees of the randomness of the EH power-supply in Figure 7(a). We set $N_s = 2$ and $M_s^{\text{low}} = 0.01$, and vary M_s^{upp} from 0.05 to 0.25, which corresponds to a larger average EH power-supply. As shown in the top-subplot of Figure 7(a), the MU's optimal total on-grid power



(a) Impact of the randomness of the EH power-supply.

(b) Impact of the number of the MUs served by the sAP.

Fig. 7: Optimal solution of Problem (OGPM) under different parameter-settings.

consumption decreases as $M_s^{\rm upp}$ increases. Such a decrease in the total on-grid power consumption essentially stems from that we rely more on the EH power-supply to power the MU's traffic offloading, which yields a decrease in the successful probability of traffic delivery, as shown in the bottom-subplot of Figure 7(a).

Impact of the Number of the Served MUs: Due to sharing the sAP's harvested energy, the number of the MUs served by the sAP (i.e., the value of N_s) will influence the optimal offloading solution of Problem (OGPM). To evaluate this impact, we plot the optimal solution versus different values of N_s in Figure 7(b). As shown in the top-subplot of Figure 7(b), the optimal total on-grid power consumption increases when N_s increases, which is due to the fact each MU is allocated a smaller amount of harvested energy when N_s increases. Correspondingly, to satisfy the MU's traffic requirement, sAP s needs to use more on-grid power to support the traffic offloading, which consequently yields an increase in the successful probability of traffic delivery, as shown in the bottom-subplot of Figure 7(b). In particular, Figure 7(b) also indicates that we need to carefully determine the numbers of the MUs served by each sAP in the multi-sAP scenario,

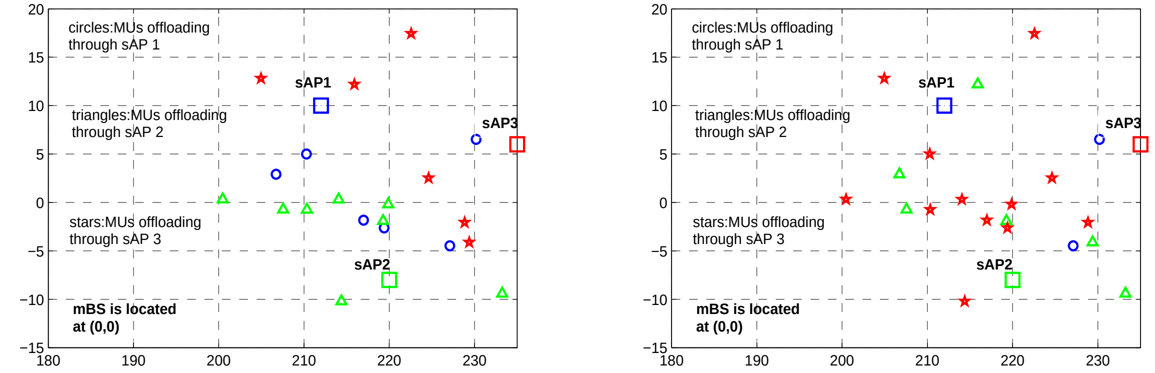
B. Numerical Results for the Case of Multiple sAPs

We next evaluate our proposed algorithms for the case of multiple sAPs, and show the performance gain of the optimal MU-selection solution of Problem (MultiMUSel). We consider a scenario that the mBS is located at the origin $(0{\rm m}, 0{\rm m})$, and three sAPs are located at $(212{\rm m}, 10{\rm m})$, $(220{\rm m}, -8{\rm m})$, and $(235{\rm m}, 6{\rm m})$. The group of MUs are randomly located within the plane whose center is $(220{\rm m}, 0{\rm m})$ and radius is 20m, i.e., the MUs are geographically closer to the sAPs than the mBS (this is a favorable condition for traffic offloading). The channel power gains and other parameters are randomly set as described before.

Illustration of the Optimal MU-Selection Solution: We first illustrate the optimal MU-selection solution of Problem (MultiMUSel), which is yielded by our proposed SelSol-Algorithm. Figure 8 shows the optimal MU-selection solution for two different cases. Specifically, Figure 8(a) shows the case of all sAPs with a homogeneous EH-capacity (i.e., $M_s^{\rm low} = 0.01$ and $M_s^{\rm upp} = 0.2, s = 1, 2, 3$). In this case, the optimal solution shows that the sAPs select different MUs to provide traffic offloading in an almost balanced way, namely, both sAP 1 and sAP 2 serve 6 MUs, and sAP 3 serves 8 MUs. Figure 8(b) shows the case of the APs with heterogenous EH-capacity. To illustrate the result clearly, we set sAP 1 with $M_1^{\rm low} = 0.01$ and $M_1^{\rm upp} = 0.2$, and sAP 2 and sAP 3 with $M_s^{\rm low} = 0.01$ and $M_2^{\rm upp} = 0.4, M_3^{\rm upp} = 0.6$, namely, sAP 3 has a much larger EH power-supply than sAP 1 and sAP 2. As a result, to fully exploit the EH power-supply and reduce the total on-grid power consumption for whole network, sAP 3 selects 12 MUs to offload traffic, as shown in Figure 8(b).

Advantage of the Optimal MU-Selection Solution: We next show the performance advantage of our SelSol-Algorithm in Figure 9. For comparison, we also show the results of a distance-based scheme in which each MU aggressively selects the sAP with the shortest distance for traffic offloading. Similar to Figure 10(b), we use the case that the sAPs are of heterogeneous EH-capacity. Figure 9(a) shows the results under different numbers of the MUs. Specifically, we fix 4 sAPs at $(212{\rm m}, 10{\rm m})$, $(220{\rm m}, -8{\rm m})$, $(235{\rm m}, 6{\rm m})$, and $(250{\rm m}, -3{\rm m})$. We vary the number of the MUs from 16 to 30, and fix each MU's traffic requirement as 40Mbps. In Figure 9(a), we mark out the relative improvement achieved by our proposed algorithm against the distance-based scheme on the top of each tested case. Figure 9(a) shows that the achieved average reward[5] gradually increases as the number of the MUs increases, since the sAPs have a larger freedom in selecting different MUs for executing the traffic offloading. Figure 9(a) validates that the average reward can be significantly increased by using our proposed SelSol-Algorithm. As shown in Figure 9(a), our SelSol-Algorithm can improve the reward up to 58.6% in comparison with the the distance-based scheme. The performance advantage not only comes from that we

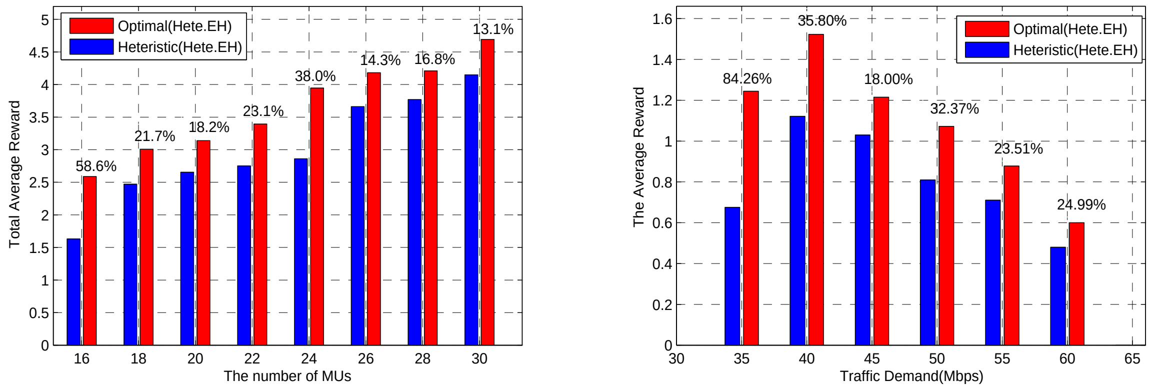
[5] Every point in Figure 9 represents the average result of 20 realizations of the MUs' geographical distributions.



(a) sAPs with homogeneous EH-capability.

(b) sAPs with heterogeneous EH-capability.

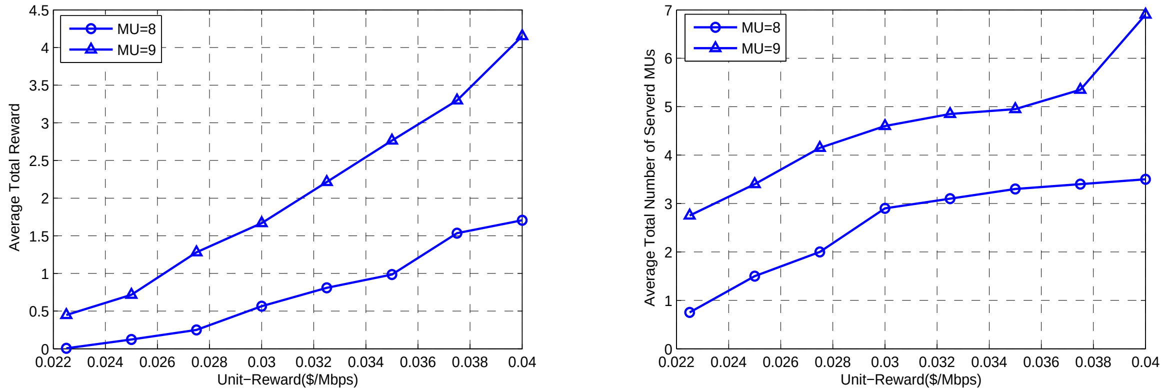
Fig. 8: Examples of the optimal MU-selection solution of Problem (MultiMUSel) under $\mu = 0.025\$/\text{Mbps}$ and $\pi = 0.002/\text{KW}$.



(a) Performance comparison under different number of MUs.

(b) Performance comparison under diff. traffic requirements.

Fig. 9: Performance advantage of the optimal MU-selection solution.



(a) The total reward versus different μ

(b) The number of served MUs versus different μ

Fig. 10: Optimal MU-selection solution under different Marginal Reward.

exploit the benefit of the DC-enabled traffic offloading which provides a flexible traffic scheduling and power allocation, but also comes from that we properly exploit the sAPs' EH power-supply by avoiding too many MUs offloading through a same sAP.

Figure 9(b) shows the results under different traffic requirements. Specifically, we fix 4 APs and 8 MUs, and vary each MU's traffic requirement from 35Mbps to 60Mbps. The results show that our proposed SelSol-Algorithm can significantly improve the average reward compared to the distance-based scheme. In particular, Figure 9(b) shows that the achieved total reward firstly increases when the traffic requirement increases, and then gradually decreases when the traffic requirement further increases beyond a threshold. This phenomenon can be explained as follows. When each MU's traffic requirement is low (i.e., $R_i^{\text{req}} < 35\text{Mbps}$), it is optimal for the sAPs to serve all MUs, and the total reward increases when each MU's R_i^{req} increases. However, when each MU's traffic requirement becomes very large (i.e., $R_i^{\text{req}} \geq 45\text{Mbps}$), the rapid increase in the total on-grid power consumption cannot be covered by the sAPs' achieved revenue to serve all MUs' traffic. As a result, the sAPs need to properly select part of the MUs to provide the traffic offloading, in order to reach a good balance between serving the MUs' traffic and reducing the total on-grid power consumption.

Impact of the Marginal Reward for Serving the MUs' Traffic: Both the marginal reward μ for successfully serving the MUs' traffic)and the marginal cost π for the total on-grid power consumption will influence the optimal number of the MUs selected by the sAPs. To show such an impact, in Figure 10, we vary μ from 0.0225\$/Mbps to 0.04\$/Mbps (while fixing $\pi = 0.02\$/\text{KW}$) and plot the corresponding results when the sAPs are of heterogenous EH-capacity. Similar to Figure 9, each point in Figure 10 corresponds to the average result of 20 random realizations of the MUs' geographical distributions. The left-subplot of Figure 10 shows that the total reward gradually increases in μ, since we can gain more for successfully serving the MUs' required traffic. Moreover, a larger μ encourages the sAPs to select more MUs (i.e., offloading traffic for more MUs), which is shown in the right-subplot.

VII. Conclusion

In this paper, we have investigated the energy-efficient DC-enabled traffic offloading through the EH-powered small cells. To reap the advantages of the DC-capability and the EH power-supply, we have proposed the joint optimization of the traffic scheduling and power allocation to minimize the total on-grid power consumption of macro and small cells. We firstly focus on the single sAP case, and have proposed an efficient layered-algorithm to compute the optimal offloading solution for each individual pair of sAP-MU. We then study the multi-sAP case and investigate how different small cells select the MUs for maximizing the total network-reward. We have also proposed an efficient algorithm to compute the optimal MU-grouping solution. Extensive numerical results have been provided to validate our proposed algorithms and the performance advantage of the proposed DC-enabled traffic offloading through the EH-powered small cells.

For our future work, we will further consider that the sAP can flexibly schedule its EH power-supply over different time slots and allocate different amounts of the EH power-supply for different MUs, and investigate the optimal design of the DC-enabled traffic offloading for improving the energy-efficiency.

References

[1] C. Rosa, K. Pedersen, and H. Wang "Dual connectivity for LTE small cell evolution: Functionality and performance aspects," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 137-143, Jun. 2016.

[2] S. Jha, K. Sivanesan, R. Vannithamby, and A. Koc, "Dual connectivity in LTE small cell networks," in *Proc. of IEEE GLOBECOM'2014 Workshops*.

[3] J. Liu, J. Liu, H. Sun, "An enhanced power control scheme for dual connectivity," in *Proc. of IEEE VTC'2014-Fall*.

[4] M. Pan, T. Lin, C. Chiu, and C. Wang, "Downlink traffic scheduling for LTE-A small cell networks with dual connectivity enhancement," *IEEE Communications Letters*, vol. 20, no. 4, pp. 796-799, Jan. 2016.

[5] Y. Wu, Y. He, L. Qian, J. Huang, and X. Shen, "Optimal resource allocations for mobile data offloading via dual-connectivity," to appear in *IEEE Transactions on Mobile Computing*, DOI:10.1109/TMC.2018.2810228, Feb. 2018.

[6] A. Mukherjee, "Macro-small cell grouping in dual connectivity LTE-B networks with non-ideal backhaul," in *Proc. of IEEE ICC'2014*.

[7] H. Wang *et. al.*, "Dual connectivity for LTE-advanced heterogeneous networks," *Wireless Networks*, vol. 22, no. 4, pp. 1315-1328, May 2016.

[8] Y. Wu, K. Guo, J. Huang, and X. Shen, "Secrecy-based energy-efficient data offloading via dual-connectivity over unlicensed spectrums," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3252-3270, Dec. 2016.

[9] R. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94-101, May 2014.

[10] M. Ismail, W. Zhuang, E. Serpedin, and K. Qaraqe, "A survey on green mobile networking: from the perspectives of network operators and mobile users," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1535-1556, Third-Quarter 2015.

[11] T. Han, N. Ansari, "On greening cellular networks via multicell cooperation," *IEEE Wireless Communications*, vol. 20, no. 1, pp. 82-89, Feb. 2013.

[12] X. Chen *et. al.*, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 627-640, Apr. 2015.

[13] S. Cai, *et. al.*, "Green 5G heterogeneous networks through dynamic small-cell operation," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1103-1114, May 2016.

[14] J. Rao *et. al.*, "Analysis of spectrum efficiency and energy efficiency of heterogeneous wireless networks with intra-inter-RAT offloading," *IEEE Trans. on Vehicluar Technology*, vol. 64, no. 7, pp. 3120-3139, Jul. 2015.

[15] H. Pervaiz *et. al.*, "Energy and spectrum efficiency trade-off for green small cell networks" in *Proc. of IEEE ICC'2015*.

[16] G. Yu, Y. Jiang, L. Xu, and G. Li, "Multi-objective energy-efficient resource allocation for multi-RAT heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2118-2127, Oct. 2015.

[17] Y. Yang, T. Quek, and L. Duan, "Backhaul-constrained small cell networks: Refunding and QoS provisioning," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 5148-5161, Sept. 2014.

[18] S. Zhang, J. Gong, S. Zhou, and Z. Niu, "How many small cells can be turned off via vertical offloading under a separation architecture?" *IEEE Transactions on Wireless Communications*, vol. 14, issue 10, pp. 5440-5453, Oct. 2015.

[19] Y. Wu, J. Chen, L. Qian, J. Huang, and X. Shen, "Energy-aware cooperative traffic offloading via device-to-device cooperations: An analytical approach," *IEEE Transactions on Mobile Computing*, vol. 16, no. 1, pp. 97-114, Jan. 2017.

[20] L. Qian, Y.J. Zhang, J. Huang, and Y. Wu, "Demand response management via real-time electricity price control in smart grids," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1268-1280, 2013.

[21] S. Zhang *et. al.*, "Energy-Aware traffic offloading for green heterogeneous networks", *IEEE Journal on Selected Areas of Communications*, vol. 34, no. 5, pp. 1116-1129, May 2016.

[22] T. Han, and N. Ansari, "ICE: Intelligent cell breathing to optimize the utilizaiton of green energy," *IEEE Communication Letters*, vol. 16, no. 6, pp. 866-869, Jun. 2012.

[23] T. Han, and N. Ansari, "Green energy aware and latency aware user associations in heterogeneous cellular networks," in *Proc. of IEEE GLOBECOM'2013 Workshops*.

[24] Y. Chia, C. Ho, and S. Sun, "Data offloading with renewable energy powered base station connected to a microgrid," in *Proc. of IEEE GlOBECOM'2014*.

[25] Z. Chang *et. al.*, "Resource allocation and data offloading for energy efficiency in wireless power transfer enabled collaborative mobile clouds," in *Proc. of IEEE INFOCOM'2015 Workshops*.

[26] J. Gong, J. Thompson, S. Zhou, and Z. Niu, "Base station sleeping and resource allocation in renewable energy powered cellular networks," *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 3801-3813, Nov. 2014.

[27] P. Yu, J. Lee, T. Quek, and Y. Hong, "Traffic offloading in heterogeneous networks with energy harvesting personal cells network throughput and energy efficiency," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1146-1161, Feb. 2016.

[28] Z. Zheng, L. Cai, R. Zhang, and X. Shen, "RNP-SA: Joint relay placement and sub-carrier allocation in wireless communication networks with sustainable energy", *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3818-3828, Oct. 2012.

[29] S. Zhou, T. Chen, W. Chen, and Z. Niu, "Outage minimization for a fading wireless link with energy harvesting transmitter and receiver," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 3, pp. 496-511, Mar. 2015.

[30] W. Li, M. L. Ku, Y. Chen, and K. Liu, "On outage probability for stochastic energy harvesting communications in fading channels," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1893-1897, Nov. 2015.

[31] A. O. Isikman, M. Yuksel, and D. Gunduz, "A low-complexity policy for outage probability minimization with an energy harvesting transmitter", *IEEE Communication Letters*, vol. 21, no. 4, pp. 917-920, Apr. 2017.

[32] P. Lee, Z. A. Eu, M. Han, and H. P. Tan, "Empirical modeling of a solar-powered energy harvesting wireless sensor node for time-slotted operation," in *Proc. of IEEE WCNC'2011*.

[33] A. Mehrabi and K. Kim, "General framework for network throughput maximization in sink-based energy harvesting wireless sensor networks," in *IEEE Transactions on Mobile Computing*, vol. 16, no. 7, pp. 1881-1896, July 1 2017.

[34] S. Guruacharya, V. Mittal, and E. Hossain, "On the battery recharge time in a stochastic energy harvesting system," available online at https://arxiv.org/pdf/1706.03183.pdf, Oct. 2017.

[35] S. Boyd, and L. Vandenberghe, "Convex Optimization," Cambridge University Press, 2004.

[36] Weisstein, Eric W. "Lambert W-Function." From MathWorld-A Wolfram Web Resource, http://mathworld.wolfram.com/LambertW-Function.html.

[37] National Instruments, "Introduction to UMTS device testing transmitter and receiver measurements for WCDMA devices," available online at http://download.ni.com/evaluation/rf/Introduction_to_UMTS_Device_Testing.pdf.

[38] J.B. Mazzola and A.W. Neebe, "Resource-constrained assignment scheduling," *Operations Research*, vol. 34, No. 4, Jul.-Aug., 1986.

[39] E. Talbi, *Metaheuristics: From design to implementation*, John Wiley & Sons Ltd, 2009.

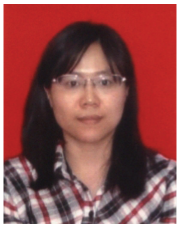
[40] D. Bertsimas, and J. Tsitsiklis, "Simulated annealing", *Statiscal Science*, vol. 8, no. 1, pp. 10-15, 1993.

[41] J.F. Martin and J.M, Sierra, "A comparison of cooling schedules for simulated annealing," Chapter of *Encyclopedia of Artificial Intelligence*, 2009, DOI:10.4018/978-1-59904-849-9.ch053.

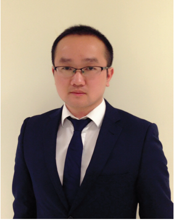
[42] B. Hajek, "Cooling schedules for optimal annealing," *Mathematics of Operations Research*, vol. 13, no. 2, pp. 311-329, May 1988.

Yuan Wu (S'08-M'10-SM'16) received the Ph.D degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He is an Associate Professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. During 2016-2017, he was with the Broadband Communications Research (BBCR) group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests focus on resource management for wireless communications and networks, and smart grid.

Xiaowei Yang is currently pursuing her M.S. degree in College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. Her research interest focuses on resource management for wireless communications and networks, and green communications.



Li Ping Qian (S'08-M'10-SM'16) received the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2010. She was with Broadband Communications Research Laboratory, University of Waterloo, from 2016 to 2017. She is currently an Associate Professor with the College of Information Engineering, Zhejiang University of Technology, China. Her research interests lie in the areas of wireless communication and networking, cognitive networks, and smart grids. Dr. Qian was a co-recipient of the IEEE Marconi Prize Paper Award in wireless communications in 2011.



Haibo Zhou received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. Since 2014, he has been a Post-Doctoral Fellow with the Broadband Communications Research Group, ECE Department, University of Waterloo. He is currently an associate professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include resource management and protocol design in cognitive radio networks and vehicular networks.

Xuemin (Sherman) Shen (IEEE M'97-SM'02-F'09) is a University Professor and the Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Canada. Dr. Shen's research focuses on wireless resource management, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is the IEEE ComSoc VP Publication, was an elected member of IEEE ComSoc Board of Governor, and the Chair of Distinguished Lecturers Selection Committee. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom'07, the Symposia Chair for IEEE ICC'10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC'08, the General Co-Chair for ACM Mobihoc'15, and the Chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/served as the Editor-in-Chief for IEEE Internet of Things Journal, and IEEE Network, a Founding Area Editor for IEEE Transactions on Wireless Communications; and an Associate Editor for IEEE Transactions on Vehicular Technology and IEEE Wireless Communications, etc. Dr. Shen received the IEEE ComSoc Education Award, the Joseph LoCicero Award for Exemplary Service to Publications, the Excellent Graduate Supervision Award in 2006, and the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.

Mohamad Khattar Awad (S'02-M'09) received the B.A.Sc. degree in electrical and computer engineering (communications option) from the University of Windsor, Windsor, ON, Canada, in 2004 and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Waterloo, ON, in 2006 and 2009, respectively. From 2004 to 2009, he was a Research Assistant with the Broadband Communications Research Group, University of Waterloo. In 2009 to 2012, he was an Assistant Professor of electrical and computer engineering with the American University of Kuwait, Kuwait City, Kuwait. Since 2012, he has been an Assistant Professor of computer engineering with Kuwait University, Kuwait City. His research interest includes wireless and wired communications, software-defined network resource allocation, wireless network resource allocation, and acoustic vector-sensor signal processing. Dr. Awad was a recipient of the Ontario Research and Development Challenge Fund Bell Scholarship in 2008 and 2009; the University of Waterloo Graduate Scholarship in 2009; a Fellowship Award from Dartmouth College, Hanover, NH, USA, in 2011; and the Kuwait University Teaching Excellence Award in 2015.