IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 9, 2016, accepted December 23, 2016, date of publication January 4, 2017, date of current version January 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2647739

Energy-Efficient WiFi Offloading and Network Management in Heterogeneous Wireless Networks

PPHUONG LUONG[1], TRI MINH NGUYEN[1], LONG BAO LE[1], (Senior Member, IEEE), NGỌC-DŨNG ĐÀO [2], (Member, IEEE), AND EKRAM HOSSAIN[3], (Fellow, IEEE)

[1] Institut National de la Recherche Scientifique, Université du Québec, Montréal, QC H5A 1K6, Canada
[2] Huawei Technologies Canada Co., Ltd., Ottawa, ON K2K 3J1, Canada
[3] University of Manitoba, Winnipeg, MB R3T 5V6, Canada

Corresponding author: L. B. Le (long.le@emt.inrs.ca)

The preliminary results of this paper have been published at IEEE WCNC'2014.

ABSTRACT We consider the joint WiFi offloading, admission control, and network management for the integrated WiFi and OFDMA-based cellular network. Specifically, we propose a quality-of-service (QoS) and mobility-aware admission control scheme that efficiently offloads macrocell traffic to WiFi and integrates a novel bandwidth *borrow-return* strategy while guaranteeing QoS requirements for users. These QoS constraints are explicitly modeled considering the throughput performance of carrier sense multiple access with collision avoidance scheme in the IEEE 802.11 WLAN and the fractional frequency reuse in the downlink cellular network. Then, we develop an analytical model to derive the blocking probabilities for calls in different service areas of a cell. Based on this analysis, we study the joint base station switching, power control, and traffic offloading problem for energy minimization under QoS constraints. We consider the time-varying call traffic and develop an algorithm to find the optimal solution of the problem. Numerical results are presented to demonstrate the performance of the proposed cross-layer resource management framework in the integrated network.

INDEX TERMS Energy efficiency, wireless LAN, power control, resource management, mobile radio mobility management.

I. INTRODUCTION

Global mobile traffic has been increasing rapidly over the last decade. Cisco has reported that the monthly global mobile data traffic will reach 30.6 exabytes by 2020 [1]. Traffic offloading to small cells (e.g., picocells and femtocells) and/or WiFi presents one of the most promising solutions to enhance the network capacity and relieve congestion for the macro cellular network [2], [3]. Another potential approach to cope with the mobile traffic growth is to allow the cellular network to share unlicensed spectrum with WiFi, which is commonly refereed to as the Long Term Evolution in unlicensed spectrum (LTE-U) technology [4]. Both WiFi offloading and LTE-U technologies aim to exploit the free-of-charge unlicensed spectrum and low-cost wireless local area network (WLAN) infrastructure to enhance the network capacity, energy and cost efficiency [5], [6]. In fact, several practical measurements indicate that 65% of mobile traffic can be offloaded to WiFi and mobile energy saving gain of 55% can be achieved [7] for the on-the-spot WiFi offloading strategy (i.e., immediate WiFi offloading), which are clearly very significant. Our current work focuses on the joint WiFi offloading, admission control, and network management for energy-efficient operations of the integrated heterogeneous cellular and WiFi network.

A. RELATED WORK

Because wireless cellular and WiFi networks have been designed and developed independently in different standards, many challenges related to the design of integrated WiFi and cellular network including network architecture design, mobility management and admission control, QoS support for mobile users, efficient interference mitigation, efficient data offloading from the cellular to WLANs, and energy-efficient network management must be addressed to realize the potential benefits of this heterogeneous wireless network. However, existing works in the literature only address some of these design issues separately.

In particular, there have been some existing works that consider the admission control design for the integrated cellular and WLANs [8]–[10]. In [8], the authors propose the so-called WLAN-first scheme to offload voice and data requests to the WLAN in the overlapped coverage area with the CDMA-based cellular network. The authors in [9] propose to combine the cutoff priority and fractional guard channel schemes to attain an enhanced admission control scheme for the integrated WLAN/cellular system. In [10], user mobility is taken into account in engineering an admission control scheme for two-tier macro-microcell networks. These existing works mainly focus on how to efficiently manage new and handoff calls as well as to optimize the radio resource utilization (WLAN and cellular spectrum) in their design. However, none of these existing admission control strategies and analytical models is suitable for the integrated WiFi and OFDMA cellular network.

A gateway-based heterogeneous network architecture supporting seamless re-direction of ongoing traffic sessions together with optimized network selection and switching design is proposed in [11]. Optimization frameworks for fair bandwidth sharing and per user throughput maximization for WiFi offloading are proposed in [12] and [13], respectively. These two papers, however, do not address the mobility management or energy efficiency optimization issues. It is reported in [7] that allowing the WiFi offloading delay of one hour or more can lead to 29% and 20% more offloaded traffic and energy saving gain, respectively compared to the on-the-spot offloading strategy. The economic benefits of delayed offloading are studied in [14] by using the Stackelberg game theory. Moreover, the work [15] studies how the successive interference cancellation (SIC) technique can impact the optimal network selection decision of mobile users when the SIC is employed at the cellular BSs or WiFi access points.

Design of energy-efficient and green wireless mobile networks has also attracted a lot of attention recently [16]. Various techniques have been proposed to lower the energy consumption of cellular wireless networks; however, design of energy-efficient traffic offloading and network management mechanisms for integrated cellular and WiFi network is under-explored in the literature. Specifically, energy-efficient hardware design is considered where reducing the energy consumption of a power amplifier (PA) in a radio BS at low traffic load is studied in [17]. In [18] and [35], the authors propose a BS switching scheme that turns off certain BSs adaptively with varying traffic loads over time and controls the zooming level of active BSs to maintain the radio coverage. The authors of [20] propose a BS switching design that can guarantee user QoS constraints and they also devise a distributed BS power control scheme to achieve further energy saving. In [21], the analysis of energy saving achieved by dynamic BS sleeping is performed. In addition, the work in [22] proposes to dynamically choose a set of active BSs from pre-determined patterns according to the time-varying traffic load to reduce network energy consumption. Another related work in [23] develops a power control scheme to achieve high SINR according to predefined sleep patterns.

B. RESEARCH CONTRIBUTIONS

It can be observed that all the existing works do not consider the joint design of WiFi offloading, mobility management and admission control, and energy-efficient network management through an appropriate BS sleeping. In addition, they usually assume over-simplified physical-layer model where important aspects of underlying radio access technologies and advanced interference mitigation techniques are not captured. This work aims to fill these important gaps where we make the following contributions.

- We propose the QoS and mobility-aware admission control (QMAC) scheme considering user mobility for WiFi offloading in integrated WLAN and OFDMA cellular network. The proposed QMAC scheme only allows slow-speed calls to be connected with the WLAN to minimize the handover overhead. This is because if high-speed calls are allowed to be connected with the WLAN then very frequent handoffs for them must be performed. In addition, each WLAN can serve slow calls located in an expanded service area to maximize the offloaded traffic from the macrocell. The QMAC scheme permits calls to be overflowed to the macrocell if they are blocked in the corresponding WLAN. We develop novel bandwidth (BW) *borrow-return* mechanism which can be integrated into the QMAC scheme to enhance the network performance.
- We describe a unified cross-layer model that characterizes the achievable throughput of the CSMA/CA protocol of WLANs and detailed channel and interference modeling for the FFR-based cellular network so that QoS guarantees for users located in WLAN-center area (WCA), WLAN-extension area (WEA), cell-center area (CCA) and cell-edge area (CEA) can be achieved. We develop an analytical model to derive call blocking probabilities in different macrocell and WLAN service areas under the proposed QMAC scheme.
- Then, we propose a novel joint macro base station (MBS) switching, traffic offloading, and power control (JMSO) design to minimize the total energy consumption considering QoS constraints for users in different network areas. Importantly, this design allows us to optimize the WLAN offloading region so that the optimum amount of macrocell traffic can be offloaded to the WLANs to minimize the energy consumption of cellular BSs. We develop an algorithm that is proved to converge to the optimal solution of the considered problem.
- Finally, we present numerical results to illustrate the performance of the admission control scheme and demonstrate the performance enhancement due to the BW *borrow-return* strategy and usefulness of the analytical model. We also illustrate the significant energy saving

that can be achieved by the proposed JMSO scheme compared to the conventional schemes.

Some preliminary results of this work have been published in [24]. However, the current manuscript makes several significant and new contributions compared to this conference version. First, the MBS switching and traffic offloading design in Section V of this journal version presents new contributions, which are not available in [24]. Second, the current journal version includes more extensive discussions of related works and detailed analysis of the considered admission control strategy and proofs of various key results. Finally, we have presented much more extensive numerical studies for both the admission control and the proposed MBS switching and traffic offloading design in this journal version where many of these numerical results are not available in the conference version [24].

The remaining of this paper is organized as follows. In Section II, we present the system model. In Section III, the admission control policy is described whose performance is analyzed in Section IV. In Section V, we present the design for joint MBS switching and offloading scheme. Numerical results are presented in Section VI followed by conclusion in Section VII.

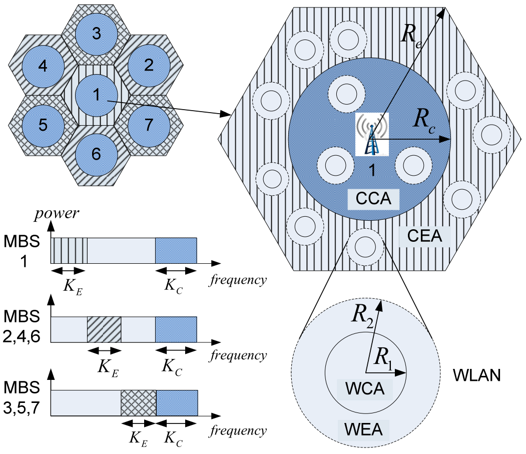


FIGURE 1. Integrated WLAN and FFR-based cellular network.

II. SYSTEM MODEL AND ASSUMPTIONS

We consider the downlink communications of an integrated WLAN and OFDMA-based cellular network. We assume that a number of WiFi Access Points (WAPs) is deployed within each macrocell of the cellular network whose macro base stations (MBSs) are located at the center of the cells. Our interest is to design an efficient and flexible traffic offloading from macrocells to WiFi. Toward this end, we assume that the service area of each WAP consists of a center coverage area (called WCA) and an extended coverage area of WLAN (called WEA) whose radii from the WAP are R_1 and R_2, respectively, which is shown in Fig. 1. Here, the center coverage area can be considered as the conventional coverage area (e.g., indoor coverage area of the WAP) while the extended coverage area is engineered to achieve desirable traffic offloading and QoS guarantees. The WAPs, which are considered in this study, can be owned and operated by the same or different wireless operators for traffic offloading purposes.

We assume that users carrying calls are categorized as either low-speed or high-speed type.[1] In this sequel, we use the terms slow and fast calls to refer to calls which are carried by high-speed and slow-speed users, respectively. Users located in the WCA and WEA, which will be called WCUs and WEUs, respectively, are only allowed to connect with the WAP if they are slow-speed ones. In addition, users located outside the WCA and WEA of any WLAN can only connect with the nearest MBS. It is worth mentioning that the WEA of each WLAN is similar to the ‘‘expanded region’’ proposed for small-cells [25]. In practice, this offloading region for a given value of R_2 can be realized by setting a suitable offset value to the range-based handoff metric so that users located in the WEA can switch their connections to the corresponding WAP [25].

We assume that the strict FFR is employed to manage the macrocell interference, which has been proposed in the LTE standard [32]. Specifically, users in each macrocell are divided into two groups, namely cell-center users (CCUs) and cell-edge users (CEUs), which are located in the CCA and CEA, respectively. In addition, the radii of the CCA and CEA are R_c and R_e, respectively. Spectrum allocation under the strict FFR is illustrated in Fig. 1 for a cluster of 3 cells. Under this FFR scheme, we allocate a common band of size K_{C} to the CCAs. The size of the band allocated to each CEA is equal to $K_{\mathrm{E}} = (B_{\mathrm{c}} - K_{\mathrm{C}})/\Delta$ where B_{c} is the total system bandwidth and Δ is the reuse factor [33]. The neighboring macro BSs (MBSs) are coordinated to ensure that their cell-edge bands are orthogonal as shown in Fig. 1. The CEUs and CCUs are restricted to access the cell-edge band and the cell-center band, respectively. The transmit power from each MBS to its intended CCUs and CEUs on a particular subchannel is assumed to be equal to $P_0 = P_{\mathrm{out}}/(k_1 + k_2)$ where P_{out} is the radio frequency (RF) output transmit power in each MBS. Hence, the number of cell-center subchannels is $k_1 = K_{\mathrm{C}}/W$ and the number of cell-edge subchannels is $k_2 = K_{\mathrm{E}}/W$ where W is the bandwidth of one subchannel.

A. QoS CONSTRAINTS IN 802.11 WLAN

Suppose that there are n users connected with a particular WAP and these users access the medium by using the distributed coordination function (DCF) based CSMA/CA protocol as specified by the 802.11 WLAN standard. We assume that all users operate in the saturated traffic regime (i.e., they always have data to transmit). Let x_{iw} be the distance from the WAP to user i and P_w be the transmit power then the signal

[1] These two types of users can be differentiated by using a suitable speed threshold, which enables us to control the required handoff signaling. Detailed designs and analysis of speed estimation, speed threshold value, and corresponding admission control signaling are outside the scope of this paper.

to noise ratio (SNR) Γ_{iw} at user i can be written as

$$\Gamma_{iw} \triangleq \frac{P_w h_{iw} \theta_{iw} L(x_{iw})}{I_{iw}} \tag{1}$$

where h_{iw} denotes the fading power gain between the WAP and its user i, θ_{iw} captures the lognormal shadowing, $L(x_{iw})$ denotes the corresponding path loss, and I_{iw} is the noise plus interference power from other WAPs. From this, we can determine the average rate for user i at distance x_i as

$$r_i = \int_0^\infty B_\mathsf{W} \log_2(1+x) f_{\Gamma_{iw}}(x) dx \tag{2}$$

where B_W is the WLAN communications bandwidth and $f_{\Gamma_{iw}}(x)$ denotes the probability density function (PDF) of Γ_{iw}. To impose the QoS constraints for users connected to an WLAN in terms of minimum average rates, we focus on the worst case where n_1 WCUs are located on the boundary of the WCA and n_2 WEUs are located on the boundary of the WEA for $n_1 + n_2 = n$. Now, we determine the throughput of one such worst user in each class. Following [26], we can compute the transmission probability τ and collision probability p_c as follows:

$$\begin{aligned} \tau &= \frac{2(1-2p_\mathrm{c})}{(W_0+1)(1-2p_\mathrm{c}) + p_\mathrm{c} W_0 (1-(2p_\mathrm{c})^{m_\mathrm{w}})} \\ p_\mathrm{c} &= 1-(1-\tau)^{n_1+n_2-1} \end{aligned} \tag{3}$$

where W_0 is the minimum contention window and m_w denotes the maximal number of backoff stages. Recall that n_i denotes the number of users in class i with the same rate r_i and corresponding successful transmission duration $T_{\mathrm{s},i}$ and collision duration $T_{\mathrm{c},i}$ where $i = 1, 2$ (i equals to 1 or 2, respectively for WCUs or WEUs). Let S_i denote the normalized throughput of a user of class i, which can be calculated as

$$S_i = \frac{p_\mathrm{s} T_{\mathrm{s},i}}{T_\mathrm{idle} + \overline{T}_\mathrm{s} + \overline{T}_\mathrm{c}} \tag{4}$$

where $p_\mathrm{s} = \tau(1-\tau)^{n_1+n_2-1}$ is the probability that a particular user transmits successfully; T_idle, $\overline{T}_\mathrm{s}$ and $\overline{T}_\mathrm{c}$ are the average idle, successful transmission and collision transmission durations, respectively. Derivations of these parameters are given in **Appendix A.** Suppose users connected with an WAP in the WCA and WEA require an average minimim rate of $S_{i,\mathrm{th}}$ (b/s) for $i = 1, 2$, respectively. Then, we have the following QoS constraints

$$S_i \times r_i \geq S_{i,\mathrm{th}}, \quad i = 1, 2. \tag{5}$$

From this, we can obtain the set of all feasible combinations (n_1, n_2) as $\Omega_\mathrm{w} = \{(n_1, n_2) : S_i r_i \geq S_{i,\mathrm{th}}, i = 1, 2\}$, which satisfy the QoS constraints in (5).

B. QoS CONSTRAINTS IN CELLULAR NETWORK

We describe the achievable signal-to-interference-plus-noise ratio (SINR) and QoS constraints for macrocell users (MUEs). User i (CCU or CEU) connected with MBS m is interfered by other MBSs. Let Q be the set of interfering MBSs of user i. Then, the SINR Γ_{im} achieved by user i associated with MBS m at distance x_{im} on one particular subchannel can be written as

$$\Gamma_{im} \triangleq \frac{P_0 h_{im} \theta_{im} L(x_{im})}{\sum_{l \in Q, l \neq m} P_0 h_{il} \theta_{il} L(x_{il}) + N_0 W} \tag{6}$$

where $h_{im}(h_{il})$ is the power channel gain from MBS m (interfering MBS $l \in Q$, $l \neq m$) to user i, which is exponentially distributed with mean μ. In addition, θ_{im} (or θ_{il}) represents the log-normal shadowing from MBS m (or MBS l) to user i, which is distributed according to a log-normal distribution $LN(A\mu_{m,dB}, A^2\sigma^2_{m,dB})$ where $A = 0.1 \ln 10$ is a scaling constant [27] and N_0 is the noise power density on each subchannel. Also, $L(x_{im})$ (or $L(x_{il})$) represents the path-loss from MBS m (or MBS l) to user i at distance x_{im} (or x_{il}). The average rate of user i (CCU or CEU) associated with MBS m at distance x_{im} can be calculated as

$$\zeta_{im} = \int_0^\infty W \log_2(1+y) f_{\Gamma_{im}}(y) dy \tag{7}$$

where W is the bandwidth of one subchannel, $f_{\Gamma_{im}}$ is the PDF of Γ_{im} that can be approximated by the lognormal distribution and determined by a numerical method as in [28] and [29]. Let $\zeta_\mathrm{min,c}$ and $\zeta_\mathrm{min,e}$ be the minimum rates achieved by the worst CCUs and CEUs, respectively ($\zeta_\mathrm{min,c}$ and $\zeta_\mathrm{min,e}$ is calculated from (7) where $x_{im} = R_c$ for CCU and $x_{im} = R_e$ for CEU). To guarantee the QoS for CCUs and CEUs, the number of subchannels that must be allocated to them should satisfy the following constraints

$$\psi_1 \geq \lceil \frac{r_\mathrm{tar,c}}{\zeta_\mathrm{min,c}} \rceil \triangleq c_1; \quad \psi_2 \geq \lceil \frac{r_\mathrm{tar,e}}{\zeta_\mathrm{min,e}} \rceil \triangleq c_2 \tag{8}$$

where $r_\mathrm{tar,c}$, $r_\mathrm{tar,e}$ are the target minimum rates of any CCUs and CEUs, respectively. The numbers of subchannels for CCUs and CEUs given in these formulas guarantee to maintain the required rates for these users regardless of their exact locations in the corresponding regions.

III. QMAC: AN ADMISSION CONTROL SCHEME FOR INTEGRATED CELLULAR AND WLANs

We propose the QoS and mobility-aware AC scheme supporting both new and handoff calls, which can be either slow-speed and high-speed ones (called slow and fast calls in the sequel). For a new or handoff slow call arriving at the WLAN areas (WCA and WEA), we assume that it always attempts to connect with the WLAN. The call is finally admitted if it is feasible to do so (i.e., the resulting numbers of calls in both WCA and WEA is in the feasible region Ω_w). Otherwise, it is blocked and overflowed to the corresponding macrocell (i.e., it attempts to connect to the MBS).

If a slow call arrives at the CCA (i.e., a new slow call or a handoff slow call or an overflowed slow call from WLAN), it will attempt to occupy the cell-center subchannels. If there are not sufficient cell-center subchannels, it will attempt to borrow the cell-edge subchannels. If there are not sufficient cell-edge subchannels to support this call, then it is blocked (dropped). If cell-center subchannels become available later (due to a cell-center call termination or leaving), the cell-center slow calls occupying the cell-edge subchannels will be shifted back to cell-center subchannels. We refer to this as the BW $borrow - return$ mechanism.

If a slow call arrives at the CEA (i.e., a new or a handoff slow call or an overflowed slow call from WLAN to the CEA), it attempts to occupy the cell-edge subchannels. If there are not sufficient cell-edge subchannels then it is blocked (dropped). The fast calls in the CCA and CEA are admitted or dropped similarly to the slow call in CCA and CEA. Recall that fast calls are not allowed to be connected with the WLAN to avoid frequent handoffs for them. Note that we do not implement the BW $borrow - return$ mechanism for calls in the CEA to avoid strong co-channel interference between users in the CEA and its neighboring CCAs.

The handover of a fast call occurs when the corresponding user crosses the cell boundaries (cell-center boundary or cell-edge boundary) for a fast call or when a slow call leaves the WEA and enters the macrocell only area and changes mobility type from the slow type to the fast one. The handover of slow call occurs as the corresponding user crosses the WLAN boundaries (WCA boundary and WEA boundary) or the macrocell boundaries (cell-center boundary or cell-edge boundary) or when a fast call from macrocell enters the WEA and changes to the slow call type.

IV. PERFORMANCE ANALYSIS OF QMAC

A. CALL MODEL AND PARAMETERS

For performance analysis, we assume an homogeneous system where there are m_1 WLANs in the CCA and m_2 WLANs in the CEA of any macrocell. We employ the isolated-cell approach for performance analysis [10]. For simplicity, we omit the macrocell index m in all notations when this does not create confusion. The new call arrivals to all areas are assumed to follow independent Poisson processes. In our model, a general call belongs to either fast or slow type with probabilities of p_f and $1-p_f$, respectively. The conversation time and sojourn time for any call in different areas are assumed to be exponentially distributed. A fast call becomes a slow call whenever it enters the WEA with probability $\gamma_{f\to s}^{c\to w}$. A slow call associated with an WLAN becomes a fast call or remains to be the slow call whenever it leaves WEA and enters the other macrocell areas with probabilities $\gamma_{s\to f}^{w\to c}$, $\gamma_{s\to s}^{w\to c}$, respectively.

B. ANALYTICAL METHOD

Recall that only slow calls can be connected to the WLAN in our proposed QMAC scheme. Let $z_1(t)$, $z_2(t)$ denote the number of slow calls located in the WCA and WEA of the considered WLAN at time t. We define the two-dimensional Markov Chain (MC) $S(t) = \{z_1(t), z_2(t) | (z_1(t), z_2(t)) \in \Omega_w\}$ and let $\bar{Z}_1, \bar{Z}_2$ represent the average values of $z_1(t)$, $z_2(t)$, respectively.

In addition, we define another MC $G(t)=\{x_s(t), x_f(t), y_s(t), y_f(t), u_s(t), u_f(t) | (x_s(t) + x_f(t))c_1 \leq k_1, (y_s(t) + y_f(t))c_1 + (u_s(t)+u_f(t))c_2 \leq k_2\}$ that describes the number of slow calls and fast calls operating on the cell-center and cell-edge bands of a macrocell defined as follows:

- $x_s(t), x_f(t)$ denote the numbers of cell-center slow and fast calls which occupy cell-center subchannels
- $y_s(t), y_f(t)$ denote the numbers of cell-center slow and fast calls which occupy cell-edge subchannels
- $u_s(t), u_f(t)$ denote the numbers of cell-edge slow and fast calls which occupy cell-edge subchannels.

Let $\bar{X}_s, \bar{X}_f, \bar{Y}_s, \bar{Y}_f, \bar{U}_s, \bar{U}_f$ denote the average values of the corresponding quantities in MC $G(t)$. In general, the two MCs $S(t)$ and $G(t)$ are coupled, which renders the exact analysis very challenging. To resolve this difficulty, we take an iterative analytical approach and analyze these two MCs in isolation. Specifically, we perform stationary analysis for these two MCs in each iteration using the handoff rates, which are updated by using the results due to the analysis performed in the previous iteration. This process is repeated until convergence. In the following, we show how to analyze the two MCs $S(t)$ and $G(t)$ and how to update the handoff arrival rates for calls in different areas by using these analytical models. Key parameters are summarized in Table 1.

1) CALCULATION OF CALL ARRIVAL RATES

For easy of exposition, call parameters related to the WCA, WEA, CCA and CEA are denoted by using notations 1W, 2W, 1C, 2C, respectively. The new slow and fast call arrival rates in the WCA, WEA, CCA, and CEA can be expressed as

$$\begin{aligned}
\lambda_{ns}^{1w} &= \lambda_d(1-p_f)\theta^{1w} \\
\lambda_{ns}^{2w} &= \lambda_m(1-p_f)\theta^{2w} \\
\lambda_{ns}^{1c} &= \lambda_m(1-p_f)(\theta^c - m_1\theta^{2w}) \\
\lambda_{nf}^{1c} &= \lambda_m p_f(\theta^c - m_1\theta^{2w}) \\
\lambda_{ns}^{2c} &= \lambda_m(1-p_f)(\theta^e - m_2\theta^{2w}) \\
\lambda_{nf}^{2c} &= \lambda_m p_f(\theta^e - m_2\theta^{2w}) \qquad (9)
\end{aligned}$$

where the call arrival rates depend on the corresponding areas, traffic densities, and parameter p_f represents the fraction of fast calls. In general, handoff events affect the system dynamics, which depend on the geographic configuration of the network [34]. To capture accurately the call handoff rates to different areas, we introduce teletraffic flow coefficients $\beta_{i,j}$ representing the average call fractions (slow or fast) which are handovered from area i to area j. In fact, $\beta_{i,j}$ can be calculated based on the perimeters of corresponding areas assuming that all moving directions are equally likely. Description of these coefficients is given in **Appendix B**. The handoff rates of slow calls or fast calls from the CCA and CEA to other areas can be expressed as

$$\begin{aligned}
\lambda_{hs}^{n\to 2c} &= \beta_{n,2c}\omega_{s,2c}\bar{U}_s \\
\lambda_{hf}^{n\to 2c} &= \beta_{n,2c}\omega_{f,2c}\bar{U}_f \\
\lambda_{hs}^{2c\to 1c} &= \beta_{2c,1c}\omega_{s,2c}\bar{U}_s \\
\lambda_{hf}^{2c\to 1c} &= \beta_{2c,1c}\omega_{f,2c}\bar{U}_f
\end{aligned}$$

TABLE 1. System parameters.

Parameter	Meaning
$m_1(m_2)$	Numbers of WLANs in CCA (CEA)
$\lambda_d)$	Traffic density of new call in WCA
λ_m	Traffic density of new call in WEA, CCA, and CEA
$\theta^c(\theta^e)$	Cell-center (cell-edge) area
$\theta^{1w}(\theta^{2w})$	WLAN center area (WLAN extension area)
μ^{-1}	Mean conversation of a call
$\omega_{f,1c}^{-1}(\omega_{f,2c}^{-1})$	Sojourn time of a fast call in CCA (CEA)
$\omega_{s,1c}^{-1}(\omega_{s,2c}^{-1})$	Sojourn time of a slow call in CCA (CEA)
$\omega_{1w}^{-1}(\omega_{2w}^{-1})$	Sojourn time of a call in WCA (WEA)
$\lambda_{ns}^{1w}(\lambda_{ns}^{2w})$	New slow calls arrival rates in WCA (WEA)
$\lambda_{ns}^{1c}(\lambda_{nf}^{1c})$	New slow (fast) calls arrival rates in CCA
$\lambda_{ns}^{2c}(\lambda_{nf}^{2c})$	New slow (fast) calls arrival rate in CEA
$\lambda_{hf}^{n\to 2c}(\lambda_{hs}^{n\to 2c})$	Handoff rate for fast (slow) calls to CEA from neighboring cells
$\lambda_{hf}^{2c\to 1c}(\lambda_{hs}^{2c\to 1c})$	Handoff rate for fast (slow) calls to CCA from the corresponding CEA
$\lambda_{hf}^{1c\to 2c}(\lambda_{hs}^{1c\to 2c})$	Handoff rate for fast (slow) calls to CEA from the corresponding CCA
$\lambda_{hs}^{1w\to 2w}$	Handoff rate for slow calls to WEA from the corresponding WCA
$\lambda_{hs}^{2w\to 1w}$	Handoff rate for calls to WCA from the corresponding WEA
$\lambda_{hs}^{2w\to 1c}(\lambda_{hs}^{2w\to 2c})$	Handoff rate for slow calls to CCA (CEA) from WEA located in the corresponding area
$\lambda_{hf}^{2w\to 1c}(\lambda_{hf}^{2w\to 2c})$	Handoff rate for fast calls to CCA (CEA) from WEA located in the corresponding area
$\lambda_{hf}^{1c\to 2w}(\lambda_{hf}^{2c\to 2w})$	Handoff rate for fast calls from CCA (CEA) to WEA due to mobility change
$\lambda_{hs}^{1c\to 2w}(\lambda_{hs}^{2c\to 2w})$	Handoff rate for slow calls from CCA (CEA) to WEA

$$\begin{aligned}
\lambda_{hs}^{2c\to 2w} &= \beta_{2c,2w}\omega_{s,2c}\bar{U}_s/m_2\\
\lambda_{hs}^{1c\to 2c} &= \beta_{1c,2c}\omega_{s,1c}(\bar{X}_s+\bar{Y}_s)\\
\lambda_{hf}^{2c\to 2w} &= \beta_{2c,2w}\omega_{f,2c}\gamma_{f\to s}^{c\to w}\bar{U}_f/m_2\\
\lambda_{hf}^{1c\to 2c} &= \beta_{1c,2c}\omega_{f,1c}(\bar{X}_f+\bar{Y}_f)\\
\lambda_{hs}^{1c\to 2w} &= \beta_{1c,2w}\omega_{s,1c}(\bar{X}_s+\bar{Y}_s)/m_1\\
\lambda_{hf}^{1c\to 2w} &= \beta_{1c,2w}\omega_{f,1c}\gamma_{f\to s}^{c\to w}(\bar{X}_f+\bar{Y}_f)/m_1. \qquad (10)
\end{aligned}$$

Note that we have included the factors m_1 and m_2 for handoff calls in the CCA and CEA since calls can enter the WLAN service areas from the macrocell equally likely. Similarly, the handoff rates of slow and fast calls from the WCA and WEA to other areas can be calculated as

$$\begin{aligned}
\lambda_{hs}^{1w\to 2w} &= \omega_{s,1w}\bar{Z}_1\\
\lambda_{hs}^{2w\to 1w} &= \beta_{2w,1w}\omega_{s,2w}\bar{Z}_2\\
\lambda_{hs}^{2w\to 1c} &= m_1\beta_{2w,1c}\omega_{s,2w}\gamma_{s\to s}^{w\to c}\bar{Z}_2\\
\lambda_{hf}^{2w\to 1c} &= m_1\beta_{2w,1c}\omega_{s,2w}\gamma_{s\to f}^{w\to c}\bar{Z}_2\\
\lambda_{hs}^{2w\to 2c} &= m_2\beta_{2w,2c}\omega_{s,2w}\gamma_{s\to s}^{w\to c}\bar{Z}_2\\
\lambda_{hf}^{2w\to 2c} &= m_2\beta_{2w,2c}\omega_{s,2w}\gamma_{s\to f}^{w\to c}\bar{Z}_2. \qquad (11)
\end{aligned}$$

2) STATIONARY ANALYSIS OF WLAN

As we defined before, MC $S(t) = \{z_1(t), z_2(t)|(z_1(t), z_2(t)) \in \Omega_w\}$ captures the states of a particular WLAN. In the following, we show how to calculate the blocking probabilities in the WCA, WEA located in CCA. The blocking probabilities in WLAN located in the CEA can be calculated similarly. For simplicity, we refer to a state with full description $(z_1(t), z_2(t))$ or simply with the state index i assuming that there is an appropriate mapping of a general state to its corresponding state index. Transition rates $q(i,j)$ from predecessor state i into state j can be expressed as

$$\begin{aligned}
q(z_1+1, z_2; z_1, z_2) &= (\mu+\omega_{s,1w})(z_1+1);\\
&\qquad (z_1+1, z_2) \in \Omega_{\mathrm{w}}\\
q(z_1, z_2+1; z_1, z_2) &= (\mu+\omega_{s,2w})(z_2+1);\\
&\qquad (z_1, z_2+1) \in \Omega_{\mathrm{w}}\\
q(z_1-1, z_2; z_1, z_2) &= \lambda_{ns}^{1w}+\lambda_{hs}^{2w\to 1w}; \quad (z_1, z_2) \in \Omega_{\mathrm{w}}\\
q(z_1, z_2-1; z_1, z_2) &= \lambda_{ns}^{2w}+\lambda_{hs}^{1c\to 2w}+\lambda_{hf}^{1c\to 2w}\\
&\quad +\lambda_{hs}^{1w\to 2w}; \quad (z_1, z_2) \in \Omega_{\mathrm{w}}.
\end{aligned}$$

Let $\pi(i)$ represent the stationary probability of state i where the states in the state space are labeled from 0 to $s_{\max}$ where each state i corresponds to a certain original state (z_1, z_2). Given the above transition rates, the stationary probabilities of all states can be determined from the set of flow balance equations and the total probability condition as follows:

$$\begin{aligned}
&\sum_{i=0}^{s_{\max}} q(i,j)\pi(i) = 0, \quad j = 0, 1, \ldots, s_{\max}\\
&\sum_{i=0}^{s_{\max}} \pi(i) = 1. \qquad (12)
\end{aligned}$$

Let us define $\lambda_{1w} = \lambda_{ns}^{1w} + \lambda_{hs}^{2w\to 1w}$ and $\lambda_{2w} = \lambda_{ns}^{2w} + \lambda_{hs}^{1c\to 2w} + \lambda_{hf}^{1c\to 2w} + \lambda_{hs}^{1w\to 2w}$ as the total arrival rates to the WCA and WEA, respectively. Then, we can calculate the blocking probabilities for slow calls in the WCA and WEA as

$$B_{1w} = \sum_{(z_1,z_2)\in S_1} \pi(z_1, z_2); \quad B_{2w} = \sum_{(z_1,z_2)\in S_2} \pi(z_1, z_2) \qquad (13)$$

where S_1 and S_2 are the sets of "blocking states" (z_1, z_2), which are defined as $S_1 = \{(z_1, z_2)|(z_1, z_2) = (\widehat{n}_1, n_2) : \widehat{n}_1 = \max\{n_1\}$ for $(n_1, n_2) \in \Omega_w\}$ and $S_2 = \{(z_1, z_2)|(z_1, z_2) =$

$(n_1, \widehat{n}_2) : \widehat{n}_2 = \max\{n_2\}$ for $(n_1, n_2) \in \Omega_w\}$. Using Little's theorem, we can compute the average number of slow calls in WCA and WEA as

$$\bar{Z}_1 = \frac{\lambda_{1w}}{\mu + \omega_{s,1w}}(1 - B_{1w}); \quad \bar{Z}_2 = \frac{\lambda_{2w}}{\mu + \omega_{s,2w}}(1 - B_{2w})$$

which are used to update the handoff arrival rates in (11).

3) STATIONARY ANALYSIS OF THE MACROCELL

We consider the isolated macrocell model represented by MC $G(t)$ that depends on the process $S(t)$ of WLANs. The MC $G(t)$ has the state space

$$G = \{(x_s, x_f, y_s, y_f, u_s, u_f) | \, 0 \le x \le k_1,$$
$$0 \le y + u \le k_2\}$$

where $x = (x_s + x_f)c_1$, $y = (y_z + y_f)c_1$ and $u = (u_s + u_f)c_2$. Since performing the stationary analysis for this 6-dimensional MC involves very high computational complexity, we propose to decompose the analysis of this MC into the analysis of cell-center and cell-edge chains and their dependence is captured through appropriate conditional probabilities. The cell-center model is represented by an MC $G_{1c}(t)$ with the state space $G_{1c} = \{(x_s, x_f); 0 \le x \le k_1\}$ and the cell-edge model is described by the MC $G_{2c}(t)$ with state space $G_{2c} = \{(y_s, y_f, u_s, u_f); 0 \le y + u \le k_2\}$.

a: STATIONARY ANALYSIS OF CELL-CENTER

Let $q_{1c}(i, j)$ denote the transition rates from predecessor state i (or $(\acute{x}_s, \acute{x}_f)$) to state j (or (x_s, x_f)) and $\pi_{1c}(i)$ denote the stationary probability of state i for MC $G_{1c}(t)$ and each state i corresponds to certain original state (x_s, x_f). We can calculate call arrival rates for slow and fast calls as

$$\begin{aligned} \lambda_{s,1c} &= \lambda_{ns}^{1c} + \lambda_{hs}^{2c\to 1c} + \lambda_{hs}^{2w\to 1c} \\ &\quad + m_1(\lambda_{1w}B_{1w} + \lambda_{2w}B_{2w}) \\ \lambda_{f,1c} &= \lambda_{nf}^{1c} + \lambda_{hf}^{2c\to 1c} + \lambda_{hf}^{2w\to 1c}. \end{aligned} \tag{14}$$

Then, conditioning on the subset of the state space of G such that $(y = 0, 0 \le y + u \le k_2)$, the cell-center probability transition rates are given in **Appendix C.** Let us define $\rho_{s,1c} = \frac{\lambda_{s,1c}}{\mu+\omega_{s,1c}}$ and $\rho_{f,1c} = \frac{\lambda_{f,1c}}{\mu+\omega_{f,1c}}$ and denote the maximum number of calls that can be served by the cell-center bands as $N_1 = \lfloor \frac{k_1}{c_1} \rfloor$. We can obtain the slow and fast call blocking probability in the product-form given no cell-center slow and fast calls occupying the cell-edge subchannels $(y = 0)$ as [10]

$$\begin{aligned} B &\triangleq P(x_s + x_f = N_1 | y = 0) = \sum_{x_s + x_f = N_1} \pi_{1c}(x_s, x_f) \\ &= \frac{\sum_{x_s=0}^{N_1} (\rho_{s,1c})^{x_s}/x_s! \times (\rho_{f,1c})^{(N_1 - x_s)}/(N_1 - x_s)!}{\sum_{x_s=0}^{N_1} (\rho_{s,1c})^{x_s}/x_s! \times \sum_{x_f=0}^{(N_1 - x_s)} (\rho_{f,1c})^{x_f}/x_f!}. \end{aligned} \tag{15}$$

This probability will be used in various derivations in the appendices.

```
Algorithm 1 Iterative Algorithm for Performance Analysis
1: Initialize the values of all handoff arrival rates to be zero.
2: Analyze the MC $S(t)$ and calculate $\bar{Z}_1, \bar{Z}_2$.
3: Analyze the MCs $G_{1c}(t)$ and $G_{2c}(t)$ and calculate $\bar{X}_s, \bar{X}_f, \bar{Y}_s, \bar{Y}_f, \bar{U}_s, \bar{U}_f$.
4: Update handoff arrival rates using (10), (11) and return to Step 2 until convergence.
```

b: STATIONARY ANALYSIS OF CELL-EDGE

The cell-edge MC $G_{2c}(t)$ has the state space $G_{2c} = \{(y_s, y_f, u_s, u_f); 0 \le y + u \le k_2\}$. Let $q_{2c}(i, j)$ denote the transition rates from predecessor state i, equivalent to $(\acute{y}_s, \acute{y}_f, \acute{u}_s, \acute{u}_f)$, to state j, equivalent to (y_s, y_f, u_s, u_f). Denote $\pi_{2c}(i)$ as the stationary probability of state i where the states in the state space G_{2c} are labeled from 0 to v and each state i corresponds to certain original state (y_s, y_f, u_s, u_f). The transition rates $q_{2c}(i, j)$ from state i to state j and the calculation of its stationary probabilities are presented in **Appendix C**.

Let us define the subset of state space G_{2c} as $\Omega = \{i$ representing states $(y_s, y_f, u_s, u_f) : y + u > \Lambda\}$ where $\Lambda = k_2 - c_2)$. The blocking probability of slow and fast call in CEA can be obtained as

$$B_{s,2c} = B_{f,2c} = P(y + u > \Lambda) = \sum_{i \in \Omega} \pi_{2c}(i). \tag{16}$$

Using the Little's theorem, the average number of slow and fast calls in CEA can be computed as

$$\bar{U}_s = \frac{\lambda_{s,2c}}{\mu + \omega_{s,2c}}(1 - B_{s,2c}); \quad \bar{U}_f = \frac{\lambda_{f,2c}}{\mu + \omega_{f,2c}}(1 - B_{f,2c})$$

which are used to update the handoff arrival rates in (10).

We now derive the blocking probabilities of slow and fast calls in CCA. Recall that a cell-center call is blocked if it cannot find sufficient subchannels, which are pre-allocated to CCA and CEA. Let us define $\Lambda_1 = k_2 - c_1$, then the blocking probability of slow and fast call in CCA can be calculated as

$$B_{s,1c} = B_{f,1c} = P(x_s + x_f = N_1, y + u > \Lambda_1) \tag{17}$$

where the derivation of $B_{s,1c} = B_{f,1c} = P(x_s + x_f = N_1, y + u > \Lambda_1)$ is given in **Appendix D**. The average number of slow and fast calls in CCA can be computed by using the Little's theorem

$$\begin{aligned} \bar{X}_s + \bar{Y}_s &= \rho_{s,1c}(1 - B_{s,1c}) \\ \bar{X}_f + \bar{Y}_f &= \rho_{f,1c}(1 - B_{f,1c}) \end{aligned} \tag{18}$$

which are used to update the handoff arrival rates in (10). Summary of the proposed analytical framework is provided in Algorithm 1. Even though we cannot prove the convergence of this iterative computation procedure, it has been widely used in the literature and we have witnessed its convergence in our extensive numerical studies.

V. MBS SWITCHING AND TRAFFIC OFFLOADING

In this section, we propose a joint MBS ON/OFF switching and offloading (JMSO) framework to engineer the green heterogeneous network by using the developed analytical model in the previous section. This framework aims to adaptively switch off as many MBSs as possible, determine the transmit powers of remaining active MBSs, and offloading region for WLANs (i.e., the radius R_2 of WEA) while still maintaining users' QoS requirements.

A. MBS ON/OFF SWITCHING MODEL

We consider a network with N MBSs. We assume that MBSs can be adaptively switched ON/OFF according to J predetermined possible switching patterns, which are also refereed to as network configurations in the following. Let $K(n)$ denote the number of switched-off MBSs out of N MBSs in network configuration n, where $n = 1, 2, \ldots, J$. We assume that a smaller network configuration index n corresponds to a smaller number of switched-off MBSs $K(n)$ without loss of generality. We define a parameter for network configuration n as $\phi(n) \triangleq \frac{K(n)}{N-K(n)}$, which is the ratio between the numbers of switched-off MBSs and the remaining active MBSs.

Examples of switching patterns for the considered hexagonal cellular network are depicted in Fig. 2 where only network configurations $n = 2, 3, 4$ are illustrated. Note that network configuration n=1 corresponds to the scenario where all MBSs are ON. For example, 2 MBSs out of every 3 MBSs are switched off in network configuration $n = 2$, which has $\phi(2)$=2. For each network configuration, active MBSs adapt their coverage accordingly by setting their transmit powers to serve traffic of the switched-off MBSs. The zoomed coverage is demonstrated by dotted lines in Fig. 2.

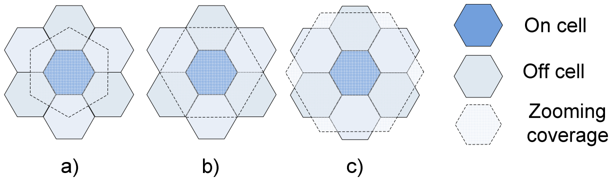


FIGURE 2. **(a) Network configuration $n = 2$ with $\phi(2) = 2$: 2 switched-off MBSs out of 3 MBSs (b) Network configuration $n = 3$ with $\phi(3) = 3$: 3 switched-off MBSs out of 4 MBSs (c) Network configuration $n = 4$ with $\phi(4) = 6$: 6 switched-off MBSs out of 7 MBSs.**

We wish to determine MBS switching patterns dynamically that can exploit the time variability of the call traffic to enhance the network energy efficiency. Toward this end, we assume that such MBS switching pattern must be determined once for each equal-size time interval (e.g., every half hour). To maintain the required users' QoS, a network configuration must be chosen according the maximum traffic load in each interval t since this corresponds to the worst user performance in terms of blocking probability.

Let $\lambda_m(t)$ and $\lambda_d(t)$ denote the maximum traffic density in the WCA and other areas (WEA, CCA, CEA) in interval t, respectively. Then, the MBS switching is designed to ensure that the required QoS can be maintained for the maximum traffic density values in each interval. Let $\theta^c(n)$, $\theta^e(n)$ denote the areas of the cell center area and cell edge area while $m_1(n)$ and $m_2(n)$ denote the number of WAPs in the CCA and number of WAPs in the CEA as network configuration n is employed. Then, we have $\theta^c(n) = (1+\phi(n))\theta^c(1)$ and $\theta^e(n) = (1+\phi(n))\theta^e(1)$ where $1+\phi(n)$ represents the increase in network coverage under network configuration n compared to network configuration one. From this we can update the new call arrival rates in CCA and CEA for network configuration n by using (9).

1) POWER CONSUMPTION OF MBS

The consumed power of an MBS typically comprises the power consumptions of a power amplifier (PA) and other parts including a radio frequency (RF) small signal transceiver module, a baseband processing engine, a DC-DC power supply, an active cooling system and an AC-DC unit [17]. In addition, the output MBS transmit power P_{out} and the PA consumed power P_{PA} are related to each other as $P_{\text{PA}} = \frac{P_{\text{out}}}{\eta_{\text{tot}}^{\text{PA}}}$ where $\eta_{\text{tot}}^{\text{PA}}$ denotes its energy efficiency considering the energy efficiency of itself as well as related components (i.e., feeder, cooling and power supply losses) [17]. The total consumed power of an MBS can be expressed as [17]

$$P_{\text{tot}} = \begin{cases} P_{\text{M}} + \dfrac{P_{\text{out}}}{\eta_{\text{tot}}^{\text{PA}}}, & 0 < P_{\text{out}} \le P_{\max} \\ P_{\text{sleep}}, & P_{\text{out}} = 0 \end{cases} \tag{19}$$

where P_{M} describes the static part of the MBS consumed power as the MBS is in an active mode, $P_{\max}$ denotes the maximum output transmit power, and P_{sleep} is the consumed power in a sleep mode (i.e., an MBS is switched off).

2) JOINT MBS SWITCHING, POWER CONTROL, AND TRAFFIC OFFLOADING PROBLEM

We are interested in determining the switching pattern n, output transmit power P_{out} for active MBSs, and radius R_2 of the WEA in each interval t so that the total network energy consumption is minimized while we can maintain the required users' QoS. For ease of exposition, we call this design problem as the joint MBS switching and offloading (JMSO) problem. Note that finding an efficient solution for the JMSO problem is challenging since the underlying design functionalities are strongly coupled. In particular, by zooming out the offloading region of the WEA, we can offload more traffic from a macrocell to WLANs which may enable us to switch off more MBSs to save power. However, the required QoSs may be violated with too large WLAN offloading region.

Without loss of generality, we consider the JMSO problem to minimize the total cellular network power consumption in one particular interval t. Applying this design for each interval of the considered time period (e.g., 48 half-hour intervals in one-day period) obviously allows us to achieve minimum total energy consumption. For brevity, we omit the

Algorithm 2 Transmit Power and Offloading Region Design

```
1: Initialize $P_{\text{out}} := P_{\max}$ and iteration index $i = 0$
2: **Repeat**
3:     Update $i := i + 1$
4:     $P_{\text{out}}(i) = \varphi P_{\text{out}}(i-1), 0 < \varphi < 1$
5:     Find optimal $R_2^*(P_{\text{out}}(i))$ in (22)
6: **Until** $B_c^{\max}(P_{\text{out}}, R_2^*) > B_c^T$
```

interval index t in all related notations. The JMSO problem in each interval t can be stated as[2]

$$\begin{aligned}\min_{n,P_{\text{out}},R_2} \quad & P_{\text{tot}} = (N - K(n))(P_{\text{M}} + \frac{P_{\text{out}}}{\eta_{\text{tot}}^{\text{PA}}}) + K(n)P_{\text{sleep}} \\ \text{s.t. } C1: \quad & B_{1c}(n, P_{\text{out}}, R_2) \le B_{\text{c}}^{\text{T}}; \\ C2: \quad & B_{2c}(n, P_{\text{out}}, R_2) \le B_{\text{c}}^{\text{T}}; \\ C3: \quad & B_{1w}(n, P_{\text{out}}, R_2) \le B_{\text{w}}^{\text{T}}; \\ C4: \quad & B_{2w}(n, P_{\text{out}}, R_2) \le B_{\text{w}}^{\text{T}} \end{aligned} \tag{20}$$

where B_{c}^{T} and B_{w}^{T} denote the maximum tolerable blocking probabilities of the macrocell blocking probabilities (i.e., B_{1c} and B_{2c}) and WLAN blocking probabilities (i.e., B_{1w} and B_{2w}), respectively; $C1$ and $C2$ represent the call blocking probability constraints of CCUs and CEUs connected with their corresponding active MBSs. In addition, $C3$ and $C4$ specify the call blocking probability constraints of WCUs and WEUs connected with the corresponding WAPs.

B. MBS POWER SETTING AND WLAN OFFLOADING SOLUTION

Note that we need to determine network configuration index n, transmit power P_{out}, and the radius of offloading region R_2 in problem (20). Since the set of possible network configurations is finite and known (i.e., $n \in \{1, \dots, J\}$), if we can determine the optimal P_{out} and R_2 for each network configuration n then we can find the optimal joint solution by comparing the objective function for all network configurations. Therefore, it is sufficient to study the problem for one particular network configuration n. In the following, we omit the dependence of the involved quantities with n if that does not create confusion.

We propose an iterative algorithm to solve this problem, which is summarized in Algorithm 2. In this algorithm, we initially set the transmit power P_{out} equal to $P_{\max}$. Then, each active MBS scales down its transmit power P_{out} by a factor $\varphi < 1$ over each iteration. For the given P_{out}, we determine the "best" radius of offloading region $R_2^*(P_{\text{out}})$ in Step 5. To clarify the operation in this step, we define $B_c^{\max}(P_{\text{out}}, R_2)$ as follows

$$B_c^{\max}(P_{\text{out}}, R_2) \triangleq \max\{B_{1c}(P_{\text{out}}, R_2), B_{2c}(P_{\text{out}}, R_2)\}. \tag{21}$$

In step 5, we determine the optimal $R_2^*(P_{\text{out}})$ as a function of transmit power P_{out} from

$$\begin{aligned}\min_{R_2} \; & B_c^{\max}(P_{\text{out}}, R_2) \\ \text{s.t. } & C3, C4. \end{aligned} \tag{22}$$

The main idea of Algorithm 2 is that we decrease the transmit powers of active MBSs until $B_c^{\max}(P_{\text{out}}, R_2^*)$ is larger the target blocking probability value B_c^T. Note that the optimization problem (22) involves only one optimization variable R_2 so its optimal solution can be found using the analytical model presented in the previous section and numerical search technique. To establish the convergence and optimality of this algorithm, we state some properties that reveal the dependence of call blocking probabilities to parameter R_2 in the following propositions.

Proposition 1: For given network configuration n and output transmit power P_{out}, the WLAN blocking probabilities $B_{1w}(n, P_{out}, R_2)$ and $B_{2w}(n, P_{out}, R_2)$ increase as the WLAN offloading region R_2 increases.

Proof: The proof is given in **Appendix F**. □

Proposition 2: For given network configuration n and output transmit power P_{out}, the macrocell blocking probabilities $B_{1c}(n, P_{out}, R_2)$ and $B_{2c}(n, P_{out}, R_2)$ first decrease and then increase as WLAN offloading region R_2 increases from R_1.

Proof: It can be seen from (14) that the total call arrival rates to CCA will decrease if the decrease in new call arrival rates due to traffic offloading to WiFi dominates the increase in the overflowed traffic rate from WiFi and vice versa. Note also that the blocking probability becomes larger with the increasing traffic arrival rate. Hence, the proposition can be proved by using the fact the total arrival rates decreases and then increases with R_2 due to the slow increase and exponential increase of the overflowed traffic rate for small R_2 (close to R_1) and large R_2, respectively. The property of $B_{2c}(n, P_{\text{out}}, R_2)$ can be proved similarly. □

The properties of blocking probabilities $B_{1c}(n, P_{\text{out}}, R_2)$ and $B_{2c}(n, P_{\text{out}}, R_2)$ stated in Proposition 2 justify why we set the WLAN offloading region R_2 as in (22), which provides the minimum macrocell blocking probability. The convergence of Algorithm 2 is stated in the following theorem.

Theorem 1: Algorithm 2 converges after a finite number of iterations for $0 < \varphi < 1$.

Proof: The proof is given in **Appendix G**. □

Algorithm 2 returns the optimal solution for problem (20), which is described in the following theorem.

Theorem 2: For $\varphi < 1$ sufficiently close to 1 (i.e., $\varphi = 1 - \epsilon$), Algorithm 2 returns the optimal solution for problem (20) under a fixed network configuration n.

Proof: It is obvious that the objective of the optimization problem (20) is minimized at the minimum value of P_{out} that still maintains its constraints. Recall that we set the

[2] We do not consider the energy-efficient design for the WLAN, which is outside the scope of this paper.

WLAN offloading region $R_2^*(P_{\text{out}})$ at the value that minimizes the macrocell blocking probability $B_c^{\max}(P_{\text{out}}, R_2)$ as in (22). In other words, $R_2^*(P_{\text{out}})$ is the best WLAN offloading region that minimizes the macrocell blocking probability $B_c^{\max}(P_{\text{out}}, R_2)$ for a given P_{out}. Since we decrease the P_{out} over iterations in Step 4 and Algorithm 2 converges according to Theorem 1, we must obtain the optimal solution for problem (20) at convergence. This completes the proof. □

Remark 1: Note that the QoS constraints for the WLAN and cellular network, given in (5) and (8), respectively, only depend on the distances of the corresponding service areas and the transmit powers in these networks. Therefore, the proposed JMSO framework can be applied as long as one knows or can estimate the time-varying traffic pattern over the considered period. Moreover, implementation of this framework requires some appropriate coordination among MBSs and WAPs, which can be performed by the mobile switching system for example.

C. ANALYSIS OF ENERGY CONSUMPTION

We analyze the energy consumption of JMSO and other schemes in this section.[3] In particular, we also consider a simplified version of JMSO in which there is no expanded offloading region, i.e., R_2 is set equal to R_1 (i.e., radius of WCA). This scheme is refereed to as the MBS switching and power setting (MSP) scheme. To obtain the solution for MSP, we still need to find the MBS transmit power using Algorithm 2; however, step 5 of this algorithm is omitted. Moreover, we are interested in the scheme where the total energy consumption when only MBS switching design is considered, which is refereed to as MBS switching only (MSO) scheme in the sequel. This means that we only need to find the maximum network configuration index, which can still maintain all QoS constraints $C1$, $C2$, $C3$ and $C4$ while the maximum output transmit power $P_{\max}$ is used and radius of offloading region is set as $R_2 = R_1$. Let $E^{\mathrm{S}}(t)$ and $E^{\mathrm{S}}_{\text{tot}}$ denote the energy consumption of scheme S in interval t and over the whole considered period, respectively. Then, we have

$$E^{\mathrm{S}}_{\text{tot}} = \sum_{t=1}^{L} E^{\mathrm{S}}(t) = \sum_{t=1}^{L} P_{\text{tot}}(n^S(t), P^S_{\text{out}}(t), R^S_2(t)) \times \Delta t \tag{23}$$

where $n^S(t)$, $P^S_{\text{out}}(t)$, and $R^S_2(t)$ denote the network configuration, output MBS transmit power, and WiFi offloading radius due to scheme S, Δt is the length of one interval, and L denotes the number of intervals in the considered period.

As a baseline scheme, we consider the case where all MBSs are active (i.e., using network configuration $n = 1$), maximum transmit power $P_{\max}$ is used and the offloading radius is set as $R_2 = R_1$. Then, the energy saving due to scheme S (i.e., JMSO, MSP, and MSO) with respect to this baseline scheme can be calculated respectively as

$$E^S_{\mathrm{s}} = 1 - \frac{\sum_{t=1}^{L} P_{\text{tot}}(n^S(t), P^S_{\text{out}}(t), R^S_2(t))}{L P_{\text{tot}}(n = 1, P_{\max}, R_2 = R_1)}. \tag{24}$$

We will study these energy saving gains in the next section where we multiply the quantities obtained in this formula with 100 to express the energy saving in %.

VI. NUMERICAL RESULTS

A. SIMULATION PARAMETERS

We evaluate the performance of the proposed admission control policy and energy-efficient radio management framework. We assume that WLAN deployment density is uniform over CCA and CEA. Therefore, the numbers of WLANs in the CCA and CEA are proportional to their corresponding areas. The number of WAPs is chosen so that the basic coverage of all WLANs (i.e., total area of all WCAs) is equal to 20% of the macrocell area unless stated otherwise. We consider a multi-cell setting with 19 macrocells and each macrocell is pre-allocated 48 subchannels over 10 MHz bandwidth. The number of cell-center subchannels k_1 and cell-edge subchannels k_2 are set to 12 for each macrocell. The radius of each macrocell is set as $R_{\mathrm{e}} = 500$ m and radius of CCA is chosen as $R_{\mathrm{c}} = 280$ m. The path loss $L(x_{iw})$ for WLAN or $L(x_{im})$ for macrocells is calculated as $L(x_{ix}) = [44.9 - 6.55 \log(h_B)] \log(x_{ix}) + 34.46 + 5.83 \log(h_B) + 23 \log(f_c/5) + \chi_{ix} L_{ix}$ where h_B is the height of WAP or MBS, f_c is the carrier frequency, χ_{ix} is the number of walls between the WAP/MBS to user i and L_{ix} is the wall loss from the WAP/MBS to user i, and x_{ix} denotes the distance between the WAP/MBS and user i.

The call arrival rate in each area is calculated by multiplying the traffic density measured in calls/min/km^2 with the area. We assume that traffic density in the WCA is k_{tr} times larger than that in other areas, i.e., $\lambda_d(t) = k_{\text{tr}} \times \lambda_m(t)$ for any time interval t. This assumption is justifiable since the WCA is relatively small but indoor traffic density has been revealed to be much higher than the outdoor traffic density. Unless stated otherwise, k_{tr} is set equal to 20 in this section. The values of the key simulation parameters are summarized in Table 2. With the simulation parameters given in this table, the numbers of subchannels required to support the minimum rates for any CCUs and CEUs are $c_1 = 2$ and $c_2 = 3$, respectively.

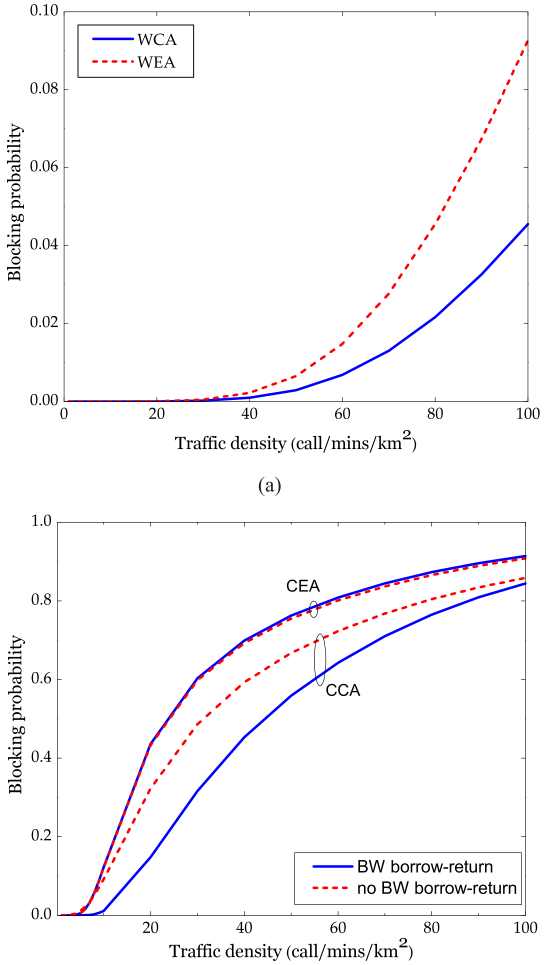
B. BLOCKING PROBABILITIES FOR CELLULAR AND WLAN USERS

In Fig. 3(a), we present the blocking probabilities of slow calls in WCA and WEA located in CCA versus the outdoor traffic density. This figure shows that the blocking probability in WCA is smaller than that in WEA and the difference between the two becomes larger as the traffic density increases. This is because the communication rate of an WCU is larger than that of an WEU. Therefore, more calls can be accepted in WCA compared to that in WEA for the

[3] Our analysis is only applied to the macrocell system where the energy consumption of all WLAN APs is not considered.

TABLE 2. Simulation parameters.

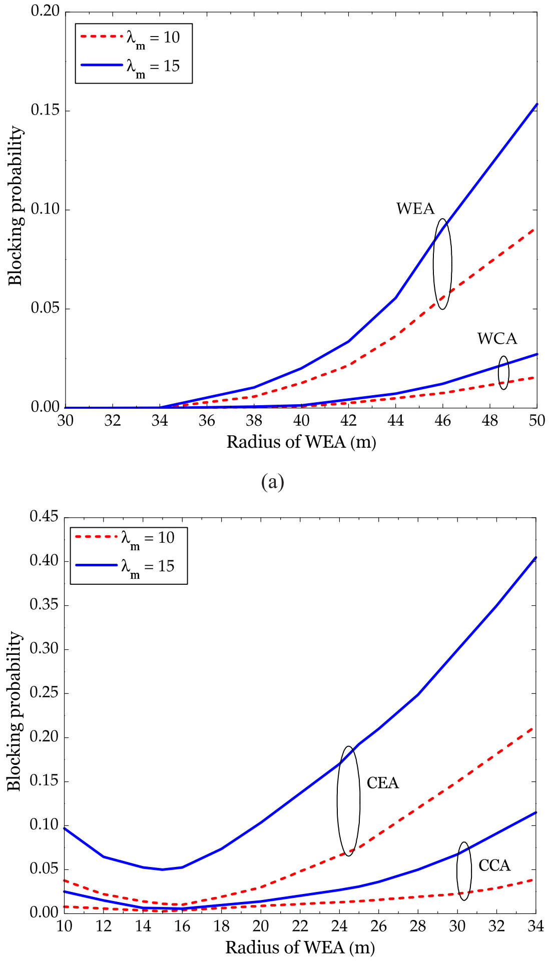
Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
P_{w}	0.1 W	$P_{\max}$	120 W	p_f	0.4	μ	1 min
h_{WAP}	10 m	h_{MBS}	25 m	$\gamma_{f\to s}^{c\to w}$	0.5	$\omega_{f,1c}$	0.3
χ_{iw}	1	χ_{im}	2	$\gamma_{s\to s}^{w\to c}$	0.35	$\omega_{f,2c}$	0.3
L_{iw}	10 dB	L_{im}	12 dB	$\gamma_{s\to f}^{w\to c}$	0.65	$\omega_{s,1c}$	0.15
f_c	2.4 GHz	f_c	2 GHz	$\eta_{\mathrm{tot}}^{\mathrm{PA}}$	0.2	$\omega_{s,2c}$	0.15
B_{w}	20 MHz	B_{c}	10 MHz	P_{M}	780 W	ω_{1w}	0.1
W_0	32	W	200 KHz	P_{sleep}	450 W	ω_{2w}	0.3
I_{iw}	-123 dBW	N_0W	-147 dBW	$SIFS, DIFS$	10 μs, 50 μs	$r_{\mathrm{tar,e}}/W$	5 b/s/Hz
σ	20 μs	σ_j	8 dB	δ	1 μs	$r_{\mathrm{tar,c}}/W$	5 b/s/Hz
L_1, L_2	2304 bytes	k_1, k_2	12	$\mathrm{S}_{1,\mathrm{th}}$	0.5 Mb/s	$\mathrm{S}_{2,\mathrm{th}}$	0.5 Mb/s



(a)

(b)

FIGURE 3. Blocking probability for calls connected with (a) WLAN in WCA and WEA, (b) MBS in CCA and CEA ($k_{\mathrm{tr}} = 20$, $R_1 = 10$m and $R_2 = 30$m).



(a)

(b)

FIGURE 4. Blocking probability for calls connected with (a) WLAN in WCA and WEA versus R_2, (b) MBS in CCA and CEA versus R_2 (traffic density $\lambda_m = 10, 15$ calls/min/km^2).

same required minimum rate. We do not show the blocking probabilities for WCUs and WEUs in the CEA since these results are similar.

Fig. 3(b) shows the blocking probabilities for calls connected with MBSs and located in CCA and CEA with and without the proposed BW *borrow-return* mechanism. Note that the blocking probabilities of the slow and fast calls are the same. It can be observed that the BW *borrow-return* mechanism can improve the blocking probability significantly. Moreover, the blocking probability of calls in CEA with and without employing the BW *borrow-return* mechanism remain almost the same. This shows the efficacy of the proposed mechanism.

In Figs. 4(a), 4(b), we illustrate the blocking probabilities for calls connected with an WAP from the WCA and WEA and for calls connected with an MBS from the CEA and CCA versus the radius R_2 of WEA, respectively. We present

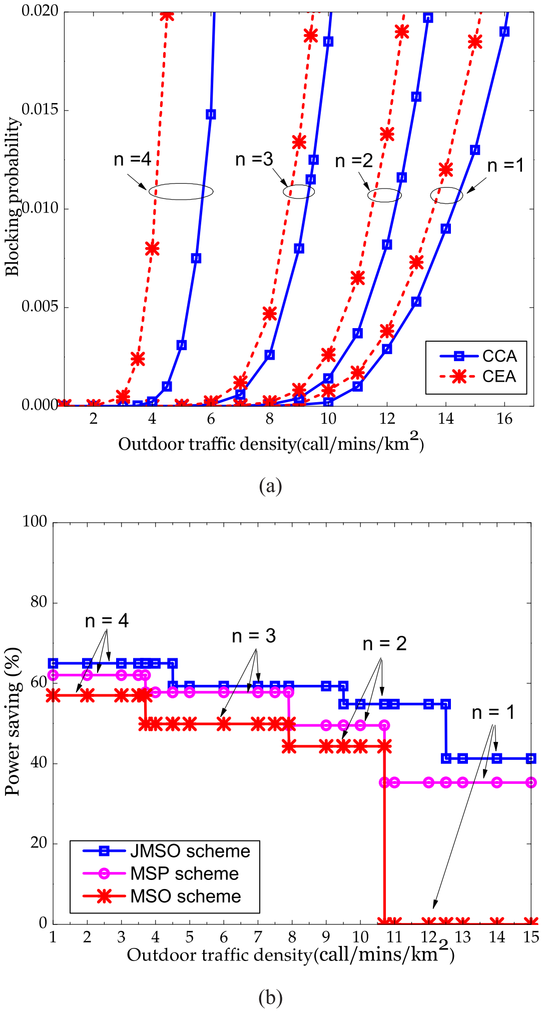


FIGURE 5. **(a) Blocking probabilities for calls connected with cellular network in CCA and CEA with varying network configurations (b) Power saving versus outdoor traffic density.**

these results for two different values of outdoor traffic density, namely, $\lambda_m = 10, 15$ (calls/min/km^2). The figures show that the blocking probabilities for calls in WCA and WEA increase quickly as the radius R_2 increases. This confirms the results stated in Proposition 1. It can be observed in Fig. 4(b) that, as R_2 increases, the blocking probabilities for calls connected with an MBS from the CCA and CEA first decreases and then increases, which validates Proposition 2. Fig. 4(b) indicates that the value of R_2 achieving minimum blocking probabilities for calls in CCA and CEA is about 15m.

In Fig. 5(a), we show the blocking probabilities for calls connected with active MBSs in CCA and CEA versus outdoor traffic density under different network configurations. In this figure, the active MBSs are assumed to set their transmit power according to the JMSO scheme. The figure suggests that higher network configurations should be employed for low traffic density in order to minimize the energy consumption. When the outdoor traffic density is sufficiently high, only network configuration $n = 1$ (i.e., all MBSs are ON) can support the required QoS.

C. ENERGY SAVING PERFORMANCE

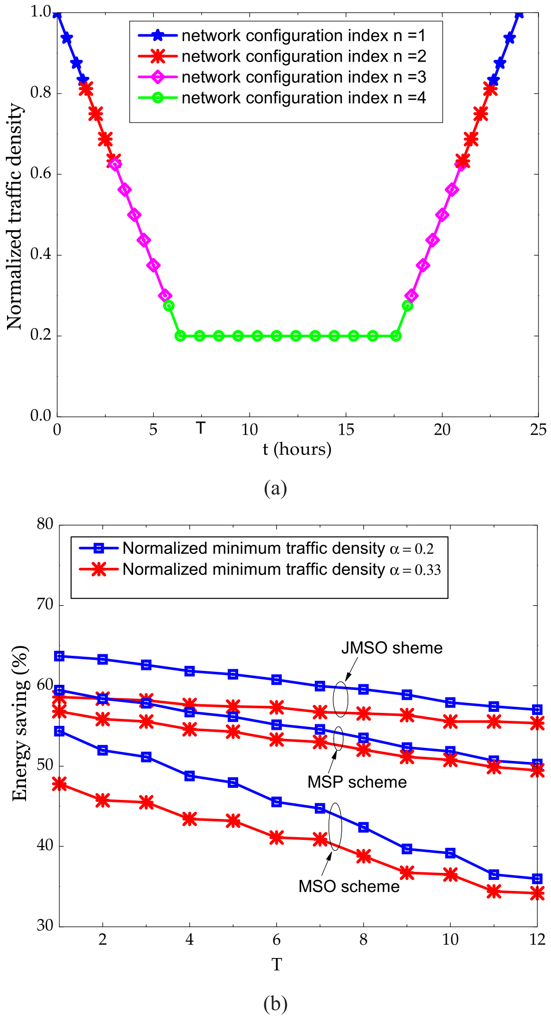
Fig. 5(b) demonstrates the energy saving achieved by the JMSO, MSP, and MSO schemes versus the outdoor traffic density, which is obtained for only one interval (i.e., from (24) with $L = 1$). The figure suggests that there are four traffic density regions in which four different network configurations should be employed. In addition, the JMSO scheme can support a larger traffic density compared to the MSP and MSO schemes under the same network configuration. Overall, the JMSO scheme achieves considerably higher energy saving compared to the other two schemes but all the schemes can achieve great energy saving under the low traffic density. In the high traffic regime where all three schemes employ network configuration one, JMSO can save up to 40% through power control while MSO scheme has no energy saving.

To illustrate the energy saving over one typical day, we consider the popular trapezoidal traffic pattern for one day as shown in Fig. 6(a). This traffic model reflects a typical practical cellular traffic pattern where there is light traffic during off-peak hours (e.g., after midnight) and heavy traffic during on-peak hours [35]. Note, however, that our proposed schemes can work with any practical traffic patterns. The symmetric trapezoidal traffic pattern is a function of time t where $t \in [0, 24]$. The peak traffic density is normalized to one at $t = 0$ and the normalized minimum traffic density during off-peak periods is equal to α $(0 \leq \alpha < 1)$. The symmetric trapezoidal traffic function $f(t)$ can be defined by parameter T $(T \in [0, 12])$ as follows:

$$f(t) = \begin{cases} 1 - t/T & 0 \leq t \leq T(1-\alpha), \\ \alpha, & T(1-\alpha) \leq t \leq 12. \end{cases} \tag{25}$$

We also assume that the 24-hour period is divided into 48 intervals, 30 minutes each, and our proposed schemes (i.e., JMSO, MSP, MSO schemes) are optimized for each interval as described in Section V. In Fig. 6(a), we show four traffic density regions where the corresponding four network configurations can be employed due to the JMSO scheme for $\alpha = 0.2$ and $T = 8$.

In Fig. 6(b), we demonstrate the energy saving achieved by the JMSO, MSP, and MSO schemes versus the values of parameters T and α of the considered traffic model. It can be observed that the energy saving reduces as T and/or α increase. This is because larger values of T reduce the length of the off-peak period, which decreases the energy saving since high network configurations are less frequently employed. Similarly, larger values of minimum traffic density α result in more frequent employment of lower network configurations, which reduces the energy saving. This figure shows that the JMSO scheme can achieve additional energy saving of up to 25% compared to that due to the MSO scheme. The MSP's energy saving is more than 5% lower than that due to the JMSO scheme for large T, which demonstrates the benefits of exploiting WiFi traffic offload in the MBS sleeping design. Moreover, the energy savings due to the JMSO and MSP schemes decline more moderately than that due to



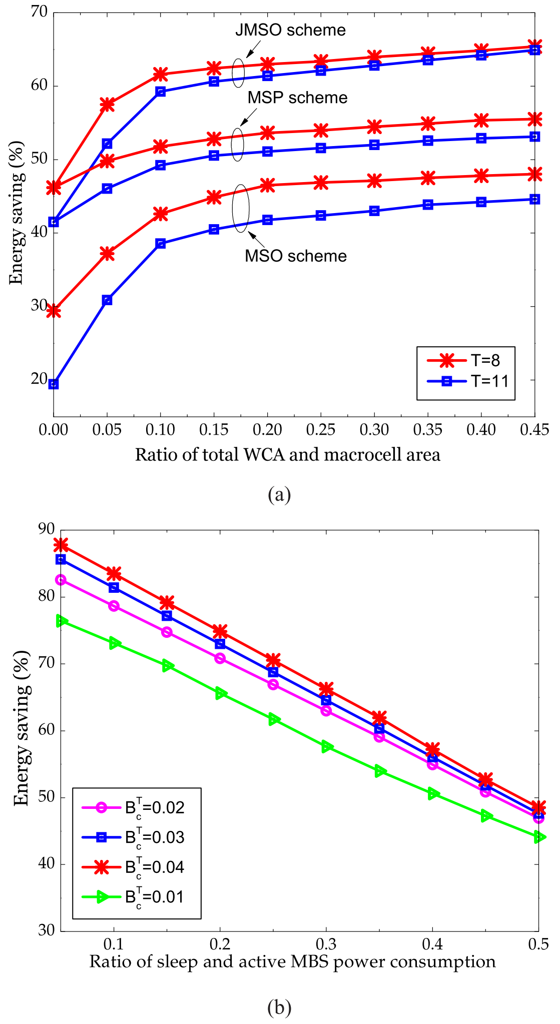
(a)

(b)

FIGURE 6. (a) Trapezoidal traffic pattern over one day with $\alpha = 0.2$, (b) Energy saving gains versus parameter T over one day.

the MSO scheme as T and α increase. For a typical traffic pattern where T is in [8, 12] and α is 0.2, we can achieve a significant energy saving of about 60% by employing JMSO scheme.

In Fig. 7(a), we illustrate the energy saving due to JMSO, MSP, and MSP schemes versus the ratio between the total area of all WLAN center areas (i.e., all WCAs) and the macrocell area for $T = 8$ or $T = 11$ and $\alpha = 0.2$. Note that a larger ratio of total WLAN center area and macrocell area implies that a larger number of WAPs are deployed in each macrocell since the radius of WLAN center area is fixed at R_1. This explains why a larger number of WAPs deployed in each macrocell result in more energy saving since more macrocell traffic can be offloaded to the WLANs. The figure also shows that the energy saving due to the JMSO scheme can be about 10% and 14% higher than those obtained by the MSP scheme for $T = 8$ and $T = 11$, respectively. This again confirms that the deployment for an expanded service area for WLANs can enable efficient offload of macrocell traffic to WiFi, which results in significant energy saving.



(a)

(b)

FIGURE 7. Energy saving gains versus (a) ratio of total WLAN center areas and macrocell area, (b) ratio of sleep and active MBS maximum power consumption.

It can also be observed that the energy saving increases slowly when the WLAN density becomes sufficiently high. In fact, WLANs can only accommodate the slow-mobility traffic and the macrocell must be configured to support the required high-mobility traffic. This imposes some constraints on the amount of offloaded traffic, which explains the saturation of the energy saving in this figure.

Finally, we plot the energy saving achieved by the JMSO scheme versus ratio of MBS sleep-mode power and active-mode maximum power in Fig. 7(b). To obtain the results in this figure, we set T equal to 8, α equal 0.2 and ratio of total WLAN center area and macrocell area is 0.4. It is evident that the energy saving decreases almost linearly with the increase of the sleep-mode power. The figure also shows that we can achieve better energy saving with larger target macrocell blocking probability as expected. In addition, for the typical scenario where MBS's sleep-mode power is 30% of the total MBS's maximum power in the active mode, we can achieve about 55% to 65% of energy saving.

VII. CONCLUSION

We have proposed a QoS and mobility-ware admission control and Wifi offloading scheme for integrated OFDMA cellular and WLANs and developed an analytical model to analyze the blocking probabilities. We have then developed a novel MBS switching, transmit power control, and traffic offloading framework to minimize the MBS energy consumption. Finally, we have studied the performance of the proposed design framework via numerical studies. Specifically, the proposed JMSO scheme achieves energy saving gains of more than 55% and about 10% compared to the conventional scheme and to the case without WiFi offloading, respectively.

APPENDIX A
PARAMETERS FOR WLAN THROUGHPUT ANALYSIS

As in [26], let us define $T_{idle} = (1 - P_{tr})\sigma$ where $P_{tr} = 1 - (1 - \tau)^{n_1+n_2}$ is the probability that there is at least one transmission in one generic slot time and σ denotes the duration of a slot time. As in [30], we have

$$\begin{aligned}\overline{T}_s &= P_s \sum_{i=1}^{2} n_i T_{s,i} \\ \overline{T}_c &= \sum_{i=2}^{n_1+n_2} \sum_{j=1}^{2} \sum_{k=1}^{n_j} \binom{n - k - \sum_{l=1}^{j-1} n_j}{i-1} T_{c,j} \\ &\quad \times \tau^i (1-\tau)^{(n_1+n_2-i)}. \qquad (26)\end{aligned}$$

According to the basic access mechanism, $T_{s,i} = T_{H,i} + T_{L,i} + SIFS + \delta + T_{ACK,i} + \delta + DIFS$ and $T_{c,i} = T_{H,i} + T_{L,i} + DIFS + \delta$ where $T_{H,i}$ is the time taken to transmit PHY and MAC header for a user in class i with transmission rate r_i; $T_{L_i,i}$ is the time to transmit a packet of size L_i of user in class i at rate r_i; $T_{ACK,i}$ is the time spent to transmit an ACK frame by a user in class i at rate r_i; $SIFS$, $DIFS$ and δ are the durations of $SIFS$, $DIFS$ and the propagation delay, respectively. For example, we have $T_{H,i} = (192 \text{ bits} + 8 \times 28 \text{ bytes})/r_i$, $T_{ACK,i} = (192 \text{ bits} + 8 \times 14 \text{ bytes})/r_i$ for the IEEE 802.11b [31].

APPENDIX B
Teletraffic Flow Coefficients

Let C_1, C_2, C_c and C_e denote the perimeter of WCA, WEA, CCA, and CEA, respectively. Then, the teletraffic flow coefficients can be expressed as

$$\beta_{n,2c} = \frac{C_e}{C_e + C_c + m_2 C_2}; \quad \beta_{2c,1c} = \frac{C_c}{C_e + C_c + m_2 C_2}$$

$$\beta_{2c,2w} = \frac{m_2 C_2}{C_e + C_c + m_2 C_2}; \quad \beta_{1c,2c} = \frac{C_c}{C_c + m_1 C_2}$$

$$\beta_{1c,2w} = \frac{m_1 C_2}{C_c + m_1 C_2}; \quad \beta_{2w,2c} = \beta_{2w,1c} = \frac{C_2}{C_2 + C_1}$$

$$\beta_{2w,1w} = \frac{C_1}{C_2 + C_1}.$$

APPENDIX C
TRANSITION RATES FOR MCs G_{1c} AND G_{2c}

Conditioning on the set of states in the state space of G_{2c} that satisfy $(y = 0, 0 \le y + u \le k_2)$, the transition rates for the cell-center MC G_{1c} can be expressed as

$$\begin{aligned}q_{1c}(x_s - 1, x_f; x_s, x_f) &= \lambda_{s,1c}; \quad c_1 \le x \le k_1 \\ q_{1c}(x_s, x_f - 1; x_s, x_f) &= \lambda_{f,1c}; \quad c_1 \le x \le k_1 \\ q_{1c}(x_s + 1, x_f; x_s, x_f) &= (\mu + \omega_{s,1c})(x_s + 1); \\ &\quad 0 \le x \le k_1 - c_1 \\ q_{1c}(x_s, x_f + 1; x_s, x_f) &= (\mu + \omega_{f,1c})(x_f + 1); \\ &\quad 0 \le x \le k_1 - c_1.\end{aligned}$$

We now describe the transition rates $q_{2c}(i, j)$ for possible transitions from state i to state j of the cell-edge MC G_{2c} in the following.

- Transition rate due to a cell-center slow (fast) call arrival which tries to occupy the cell-edge subchannels: According to our admission control scheme, the cell-center slow (fast) calls only try to occupy the cell-edge subchannels if cell-center subchannels are fully used. It is clear that, for the case with $y = 0$, as a cell-center slow or fast call arrives, it will try to occupy the cell-edge subchannels if it is blocked in cell-center with conditional blocking probability B calculated before. For cases where $y_s + y_f \ge 1$, all cell-center subchannels are already fully utilized; therefore, any slow (fast) call arriving at the CCA will always try to occupy the cell-edge subchannels. Hence, we obtain the transition rate as
$$\begin{aligned}&q_{2c}(y_s - 1, y_f, u_s, u_f; y_s, y_f, u_s, u_f) \\ &\quad = \begin{cases} B\lambda_{s,1c}\,; \ y = c_1, c_1 \le y + u \le k_2 \\ \lambda_{s,1c}\,; \ y > c_1, c_1 < y + u \le k_2 \end{cases} \\ &q_{2c}(y_s, y_f - 1, u_s, u_f; y_s, y_f, u_s, u_f) \\ &\quad = \begin{cases} B\lambda_{f,1c}\,; \ y = c_1, c_1 \le y + u \le k_2 \\ \lambda_{f,1c}\,; \ y > c_1, c_1 < y + u \le k_2. \end{cases}\end{aligned}$$
- Transition rate due to cell-edge slow (or fast) call arrival given state i $(y_s, y_f, u_s - 1, u_f)$ (or $(y_s, y_f, u_s, u_f - 1)$) to state j (y_s, y_f, u_s, u_f): Let $\lambda_{s,2c} = \lambda_{ns}^{2c} + \lambda_{hs}^{1c\to 2c} + \lambda_{hs}^{n\to 2c} + \lambda_{hs}^{2w\to 2c} + m_2(\lambda_{1w}B_{1w} + \lambda_{2w}B_{2w})$ and $\lambda_{f,2c} = \lambda_{nf}^{2c} + \lambda_{hf}^{1c\to 2c} + \lambda_{hf}^{n\to 2c} + \lambda_{hf}^{2w\to 2c}$. Then, we have
$$\begin{aligned}q_{2c}(i, j) &= \lambda_{s,2c}; \quad c_2 \le y + u \le k_2 \\ q_{2c}(i, j) &= \lambda_{f,2c}; \quad c_2 \le y + u \le k_2.\end{aligned}$$
- Transition rate due to the departure of cell-center slow (or fast) call which is occupying the cell-edge subchannels given the state i $(y_s + 1, y_f, u_s, u_f)$ (or $(y_s, y_f + 1, u_s, u_f)$) to state j (y_s, y_f, u_s, u_f): The departure can be due to a call completion or a shifting back to the cell-center subchannels due to the cell-center calls leaving. Thus, we have
$$\begin{aligned}q_{2c}(i, j) &= (\mu + \omega_{s,1c})(y_s + 1) + N_1(\mu + \omega_{s,1c}); \\ &\quad 0 \le y + u \le (k_2 - c_1)\end{aligned}$$

$$\begin{aligned} q_{2c}(i,j) &= (y_f + 1 + N_1)(\mu + \omega_{f,1c}); \\ &\quad 0 \le y + u \le (k_2 - c_1). \end{aligned}$$

- Transition rate due to a departure of cell-edge slow (or fast) call given state i $(y_s, y_f, u_s + 1, u_f)$ (or $(y_s, y_f, u_s + 1, u_f)$) to state j (y_s, y_f, u_s, u_f):

$$\begin{aligned} q_{2c}(i,j) &= (u_s + 1)(\mu + \omega_{s,2c}); \\ &\quad 0 \le y + u \le (k_2 - c_2) \\ q_{2c}(i,j) &= (u_f + 1)(\mu + \omega_{f,2c}); \\ &\quad 0 \le y + u \le (k_2 - c_2). \end{aligned}$$

From these transition rates, we can calculate the stationary probabilities of this MC similarly to (12).

APPENDIX D
DERIVATION OF $B_{s,1c} = B_{f,1c} = P(x_s + x_f = N_1, y + u > \Lambda_1)$

We have

$$\begin{aligned} &P(x_s + x_f = N_1, y + u > \Lambda_1) \\ &\quad = P(x_s + x_f = N_1, y = 0, u > \Lambda_1) \\ &\qquad + P(x_s + x_f = N_1, y > 0, y + u > \Lambda_1) \end{aligned}$$

where

$$\begin{aligned} &P(x_s + x_f = N_1, y = 0, u > \Lambda_1) \\ &\quad = P(x_s + x_f = N_1 | y = 0, u > \Lambda_1) P(y = 0, u > \Lambda_1) \\ &\quad = P(x_s + x_f = N_1 | y = 0) P(y = 0, u > \Lambda_1) \qquad (27) \end{aligned}$$

and

$$\begin{aligned} &P(x_s + x_f = N_1, y > 0, y + u > \Lambda_1) \\ &\quad = P(x_s + x_f = N_1 | y > 0, y + u > \Lambda_1) \\ &\qquad \times P(y > 0, y + u > \Lambda_1) \\ &\quad = P(x_s + x_f = N_1 | y > 0) P(y > 0, y + u > \Lambda_1). \quad (28) \end{aligned}$$

The proof of conditional independence in equations (27) and (28) is provided in **Appendix E**. Let us now define the subset of state space G_{2c} as $\Omega_1 = \{\underline{i} = (y_s, y_f, u_s, u_f) : y = 0, u > \Lambda_1\}$ and $\Omega_2 = \{\underline{i} = (y_s, y_f, u_s, u_f) : y > 0, y + u > \Lambda_1\}$. The probabilities $P(y = 0, u > \Lambda_1)$ and $P(y > 0, y + u > \Lambda_1)$ can be computed from the stationary probabilities of cell-edge MC model as follows:

$$\begin{aligned} P(y = 0, u > \Lambda_1) &= \sum_{\underline{i} \in \Omega_1} \pi_{2c}(\underline{i}) \\ P(y > 0, y + u > \Lambda_1) &= \sum_{\underline{i} \in \Omega_2} \pi_{2c}(\underline{i}). \qquad (29) \end{aligned}$$

We already have $P(x_s + x_f = N_1 | y = 0) = B$ and $P(x_s + x_f = N_1 | y > 0) = 1$. Therefore, we can get the slow and fast blocking probabilities $B_{s,1c}$ and $B_{f,1c}$.

APPENDIX E
PROOF OF CONDITIONAL INDEPENDENCE

Let us define $x_1(t) = x_s(t) + x_f(t) + y_s(t) + y_f(t)$ and $x_2(t) = u_s(t) + u_f(t)$ as the total number of cell-center calls and cell-edge calls associated with MBS at time t. Due to the *shift-back* mechanism adopted in the QMAC scheme, we have $x_s(t) + x_f(t) = \min\{x_1(t), N_1\}$ and $y_s(t) + y_f(t) = x_1(t) - (x_s(t) + x_f(t))$. Let $N_2 = \lfloor k_2/c_1 \rfloor$ and $N_3 = \lfloor k_2/c_2 \rfloor$. Thus, the process $\{(x_1(t), x_2(t))\}$ has the product-form stationary distribution. This process has the state space $\Psi = \{\underline{x} : 0 \le x_1 \le N_1 + N_2, 0 \le x_2 \le N_3, 0 \le x_1 + x_2 \le N_1 + \max\{N_2, N_3\}\}$ where vector $\underline{x} = (x_1, x_2)$. We have

$$\pi(x_1, x_2) = \frac{1}{F} \phi_{x_1}(x_1) \phi_{x_2}(x_2)$$

where F is the normalized constant. We start the proof for equation (27) by using the following results:

$$\begin{aligned} &P(x_s + x_f = N_1, y = 0, u > \Lambda_1) \\ &\quad = \frac{1}{F} \phi_{x_1}(N_1) \sum_{\lfloor \frac{\Lambda_1}{c_2} \rfloor < x_2 < N_3} \phi_{x_2}(x_2) \\ &P(y = 0, u > \Lambda_1) \\ &\quad = \frac{1}{F} \sum_{0 \le x_1 \le N_1} \phi_{x_1}(x_1) \sum_{\lfloor \frac{\Lambda_1}{c_2} \rfloor < x_2 < N_3} \phi_{x_2}(x_2). \end{aligned}$$

Therefore, we have

$$\begin{aligned} &P(x_s + x_f = N_1 | y = 0, u > \Lambda_1) \\ &\quad = \frac{\phi_{x_1}(N_1)}{\sum_{0 \le x_1 \le N_1} \phi_{x_1}(x_1)}. \end{aligned}$$

On the other hand, we also have

$$\begin{aligned} &P(x_s + x_f = N_1, y = 0) \\ &\quad = \frac{1}{F} \phi_{x_1}(N_1) \sum_{0 \le x_2 \le N_3} \phi_{x_2}(x_2) \\ &P(y = 0) = \frac{1}{F} \sum_{0 \le x_1 \le N_1} \phi_{x_1}(x_1) \sum_{0 \le x_2 \le N_3} \phi_{x_2}(x_2). \end{aligned}$$

Hence,

$$P(x_s + x_f = N_1 | y = 0) = \frac{\phi_{x_1}(N_1)}{\sum_{0 \le x_1 \le N_1} \phi_{x_1}(x_1)}.$$

It follows that equation (27) holds. Let $N_4 = \lfloor \frac{\Lambda_1 - (x_1 - N_1) c_1}{c_2} \rfloor$. We start the proof of equation (28) with the following results

$$\begin{aligned} &P(x_s + x_f = N_1, y > 0, y + u > \Lambda_1) \\ &\quad = \frac{1}{F} \sum_{N_1 < x_1 \le N_1 + N_2} \phi_{x_1}(x_1) \sum_{N_4 < x_2 \le N_3} \phi_{x_2}(x_2). \end{aligned}$$

Due to the *shift-back* mechanism of the QMAC, we have

$$\begin{aligned} &P(y > 0, y + u > \Lambda_1) \\ &\quad = \frac{1}{F} \sum_{N_1 < x_1 \le N_1 + N_2} \phi_{x_1}(x_1) \sum_{N_4 < x_2 \le N_3} \phi_{x_2}(x_2). \end{aligned}$$

Therefore, we have

$$P(x_s + x_f = N_1 | y > 0, y + u > \Lambda_1) = \frac{P(x_s + x_f = N_1, y > 0, y + u > \Lambda_1)}{P(y > 0, y + u > \Lambda_1)} = 1.$$

Let us define $N_5 = \lfloor \frac{k_2 - (x_1 - N_1)c_1}{c_2} \rfloor$. Similarly, we have

$$P(x_s + x_f = N_1, y > 0) = \frac{1}{F} \sum_{N_1 < x_1 \le N_1 + N_2} \phi_{x_1}(x_1) \sum_{0 \le x_2 \le N_5} \phi_{x_2}(x_2).$$

and

$$P(y > 0) = \frac{1}{F} \sum_{N_c < x_1 \le N_c + N_1} \phi_{x_1}(x_1) \sum_{0 \le x_2 \le N_5} \phi_{x_2}(x_2).$$

From that, we have $P(x_s + x_f = N_1 | y > 0) = 1$. Therefore, equation (28) holds.

APPENDIX F
PROOF OF PROPOSITION 1

Consider two different values of WLAN offloading radius $R'_2 > R_2$. Then, the SNRs for user i located at the boundary of WEA and connected with an WAP for these offloading radii R'_2 and R_2 satisfy

$$\Gamma_{iw}(R'_2) = \frac{P_w h_{iw} \theta_{iw} L(R'_2)}{P_{\text{noise}}} < \frac{P_w h_{iw} \theta_{iw} L(R_2)}{P_{\text{noise}}} = \Gamma_{iw}(R_2).$$

This means that the SNR of any user located at the boundary of WEA decreases, which results in the decrease of the average rate of the user according to (2) as R_2 increases. Consequently, the feasible region for (n_1, n_2) Ω_{w} (i.e., Ω_{w} contains all possible (n_1, n_2) whose QoS requirements in (5) can be supported) shrinks as R_2 increases. In addition, the traffic arrival rates to the WEA increases as R_2 increases. Therefore, the call blocking probabilities $B_{1\text{w}}(n, P_{\text{out}}, R_2)$ and $B_{2\text{w}}(n, P_{\text{out}}, R_2)$ in the WCA and WEA increase with R_2. Therefore, we have completed the proof of the proposition.

APPENDIX G
PROOF OF THEOREM 1

First, we prove that the macrocell blocking probabilities of all users associated to any active MBSs increase when these MBSs decrease their transmit power P_{out}. Let us consider a particular user j connecting to an active MBS m and let $Q(n)$ be the set of active MBSs interfering user j. In the following, we compare the SINRs Γ_{jm} achieved by user j associated with active MBS m in two iterations i and $i+1$. We have

$$\Gamma_{jm}(i+1) = \frac{\varphi P_0(i) h_{jm} \theta_{jm} L(x_{jm})}{\sum_{l \in Q(n), l \neq m} \varphi P_0(i) h_{jl} \theta_{jl} L(x_{jl}) + \delta^2} = \frac{P_0(i) h_{jm} \theta_{jm} L(x_{jm})}{\sum_{l \in Q(n), l \neq m} P_0(i) h_{jl} \theta_{jl} L(x_{jl}) + \frac{\delta^2}{\varphi}} < \Gamma_{jm}(i).$$

This means that the SINR of any users connected with an active MBS decreases as all active MBSs reduce their transmit powers by the same factor $\varphi < 1$. This leads to the decrease in the average rate of users connected with active MBSs as they reduce their transmit powers by referring to (7). Therefore, the minimum numbers of subchannels required to meet the minimum rates in (8) increase. As a result, the macrocell blocking probabilities increase as active MBSs reduce their transmit powers since the total numbers of available subchannels for CCA and CEA in each active macrocell are fixed at k_1 and k_2, respectively.

We have proved that the blocking probabilities of all users associated with active MBSs increase when these MBSs decrease their transmit powers P_{out} by the same factor $\varphi < 1$. From this we have

$$B_c^{\max}(P_{\text{out}}(i+1), R_2^*(P_{\text{out}}(i+1))) \ge B_c^{\max}(P_{\text{out}}(i), R_2^*(P_{\text{out}}(i+1))).$$

In addition, we have

$$B_c^{\max}(P_{\text{out}}(i), R_2^*(P_{\text{out}}(i+1))) \ge B_c^{\max}(P_{\text{out}}(i), R_2^*(P_{\text{out}}(i)))$$

since $R_2^*(P_{\text{out}}(i))$ achieves minimum value of $B_c^{\max}$ at $P_{\text{out}}(i)$. These two inequalities imply that

$$B_c^{\max}(P_{\text{out}}(i+1), R_2^*(P_{\text{out}}(i+1))) \ge B_c^{\max}(P_{\text{out}}(i), R_2^*(P_{\text{out}}(i))).$$

Hence, the minimum blocking probabilities achieved at $R_2^*(P_{\text{out}})$ also increase as the MBSs reduce their transmit powers. Thus, we can conclude that Algorithm 2 must terminate after a finite number of iterations since the macrocell blocking probabilities will exceed the target threshold B_c^T otherwise. This completes the proof of the theorem.

REFERENCES

[1] (Feb. 2016). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019*. [Online]. Available: www.cisco.com/c/en/us/solutions/collateral/service-provider/visualnetworking-index-vni/white_paper_c11-520862.pdf

[2] F. Rebecchi, M. D. de Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti, "Data offloading techniques in cellular networks: A survey," *IEEE Commun. Surveys Tut.*, vol. 17, no. 2, pp. 580–603, 2nd Quart., 2015.

[3] M. Bennis, M. Simsek, A. Czylwik, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *IEEE Comm. Mag.*, vol. 51, no. 6, pp. 44–50, Jun. 2013.

[4] H. Zhang, X. Chu, W. Guo, and S. Wang, "Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 158–164, Mar. 2015.

[5] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. ACM MobiSys* San Francisco, CA, USA, Jun. 2010, pp. 209–222.

[6] B. Halvarsson and O. Zee, " Novel Wi-Fi—LTE real-time traffic steering—first field measurement results," in *Proc. IEEE Veh. Technol. Conf. (IEEE VTC-Fall)*, Sep. 2014, pp. 1–5.

[7] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 536–550, Apr. 2013.

[8] W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the WLAN-first scheme in cellular/WLAN interworking," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1932–1952, May 2007.

[9] S.-N. Enrique, A. H. Mohsenian-Rad, and W. S. W. Vincent, "Connection admission control for multiservice integrated cellular/WLAN system," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3789–3800, Nov. 2008.

[10] K. Maheshwari and A. Kumar, "Performance analysis of microcellization for supporting two mobility classes in cellular wireless networks," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 321–333, Mar. 2000.

[11] R. Mahindra, H. Viswanathan, K. Sundaresan, M. Y. Arslan, and S. Rangarajan, "A practical traffic management system for integrated LTE-WiFi networks," in *Proc. ACM MobiCom* Sep. 2014, pp. 189–200.

[12] Y. Im, C. Joe-Wong, S. Ha, S. Sen, T, T. Kwon, and M. Chiang, "AMUSE: Empowering users for cost-aware offloading with throughput-delay tradeoffs," *IEEE Trans. Mobile Comput.*, vol. 15, no. 5, pp. 1062–1076, May 2016.

[13] B. H. Jung, N.-O. Song, and D. K. Sung, "A network-assisted user-centric WiFi-Offloading model for maximizing per-user throughput in a heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1940–1945, May 2014.

[14] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1540–1554, Mar. 2014.

[15] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile data offloading through a third-party WiFi access point: An operator's perspective," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5340–5351, Oct. 2014.

[16] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, Nov. 2011.

[17] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[18] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

[19] Z. Niu, "TANGO: Traffic-aware network planning and green operation," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 25–29, Oct. 2011.

[20] L. B. Le, "QoS-aware BS switching and cell zooming design for OFDMA green cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 1544–1549.

[21] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, Dresden, Germany, Jun. 2009, pp. 1–5.

[22] L. Chiaraviglio, D. Ciullo, M. Meo, and M. A. Marsan, "Energy-efficient management of UMTS access networks," in *Proc. 21st Int. Teletraffic Congr. (ITC)*, Paris, France, Sep. 2009, pp. 1–8.

[23] Z. Pan and S. Shimamoto, "Cell sizing based energy optimization in joint macro-femto deployments via sleep activation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 4765–4770.

[24] P. Luong, T. M. Nguyen, B. L. Le, and N.-D. Dào, "Admission control design for integrated WLAN and OFDMA-based cellular networks," in *Proc. IEEE Wireless Commun. Net. Conf. (WCNC)*, Apr. 2014, pp. 1386–1391.

[25] D. L Opez-Pérez, X. Chu, and I. Guvenc, "On the expanded region of picocells in heterogeneous networks," *IEEE J. Sel. Topics Sig. Process.*, vol. 6, no. 3, pp. 281–294, Jun. 2012.

[26] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[27] V. Chandrasekhar and J. G. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[28] L. B. Le, D. Niyato, E. Hossain, D. I. Kim, and D. T. Hoang, "QoS-aware and energy-eficient resource management in OFDMA femtocells," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 180–194, Jan. 2013.

[29] N. B. Mehta, J. Wu, A. F. Molisch, and J. Zhang, "Approximating a sum of random variables with a lognormal," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2690–2699, Jul. 2007.

[30] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions," *IEEE/ACM Trans. Netw.*, vol. 15, no. 1, pp. 159–172, Feb. 2007.

[31] D.-Y. Yang, T.-J. Lee, K. Jang, J.-B. Chang, and S. Choi, "Performance enhancement of multirate IEEE 802.11 WLANs with geographically scattered stations," *IEEE Trans. Mobile Comput.*, vol. 5, no. 7, pp. 906–919, Jul. 2006.

[32] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 74–81, Apr. 2009.

[33] T. D. Novlan, R. K. Ganti, A. Ghosh, and J. G. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4294–4305, Dec. 2011.

[34] S. Rappaport Stephen and L.-R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis," *Proc. IEEE*, vol. 82, no. 9, pp. 1383–1397, Sep. 1994.

[35] Z. Niu, "TANGO: Traffic-aware network planning and green operation," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 25–29, Oct. 2011.

PHUONG LUONG received the B.Eng. degree in electrical engineering from the Hanoi University of Science and Technology, Vietnam, in 2009, and the M.E. degree in electrical engineering from Kyung Hee University, Yongin, South Korea, in 2012. Her current research interests are in the areas of wireless communications and networking.

TRI MINH NGUYEN received the B.Eng. degree in electrical engineering from the Ho Chi Minh City University of Technology, Vietnam, in 2009, and the M.E. degree in electrical engineering from Kyung Hee University, Yongin, South Korea, in 2012. His current research interests include interference management in heterogeneous networks and MIMO communications.

LONG BAO LE (S'04–M'07–SM'12) received the B.Eng. degree in electrical engineering from the Ho Chi Minh City University of Technology, Vietnam, in 1999, the M.Eng. degree in telecommunications from the Asian Institute of Technology, Thailand, in 2002, and the Ph.D. degree in electrical engineering from the University of Manitoba, Canada, in 2007. He was a Post-Doctoral Researcher with the Massachusetts Institute of Technology (2008-2010) and the University of Waterloo (2007-2008). Since 2010, he has been with the Institut National de la Recherche Scientifique, Université du Québec, Montréal, QC, Canada, where he is currently an Associate Professor. He is a co-author of the book *R*adio Resource Management in Multi-Tier Cellular Wireless Networks (Wiley, 2013). His current research interests include smart grids, cognitive radio, radio resource management, network control and optimization, and emerging enabling technologies for 5G wireless systems. He is a member of the Editorial Board of the IEEE Transactions on Wireless Communications and the IEEE Communications Surveys and Tutorials. He has served as a Technical Program Committee Chair and a Co-Chair for several IEEE conferences including IEEE WCNC, IEEE VTC, and IEEE PIMRC.

NGỌC-DŨNG ĐÀO (S'02–M'07) received the Ph.D. degree from the University of Alberta, Canada, in 2006. He was with Siemens Communications Networks in Vietnam from 1995 to 2000, Toshiba Research Europe, U.K., from 2007 to 2010. He is currently a Senior Research Engineer with Huawei Technologies Canada Co. Ltd. His recent research interests include SDN and NFV for radio access networks, information-centric networking, heterogeneous and dense networks, and reliable video communications for 5G mobile networks. He is an Editor of the IEEE Transactions on Vehicular Technology and the IEEE Communications Surveys and Tutorials, an Associate Technical Editor of the *I*EEE Communications Magazine, and a Guest Editor of 5G Special Issue of *E*URASIP Journal on Wireless Communications and Networking.

EKRAM HOSSAIN (F'15) received the Ph.D. degree in electrical engineering from the University of Victoria, Canada, in 2001. He has been a Professor with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada, since 2010. He has authored/edited several books in these areas. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He is currently a member (Class of 2016) of the College of the Royal Society of Canada. He is a member of the IEEE Press Editorial Board. He serves as the Editor-in-Chief of the IEEE Communications Surveys and Tutorials and an Editor of the IEEE Wireless Communications. He has received several research awards including the IEEE Communications Society Transmission, Access, and Optical Systems Technical Committee's Best Paper Award in IEEE Globecom in 2015, the University of Manitoba Merit Award for Research and Scholarly Activities in 2010, 2014, and 2015, the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 Best Paper Award. He served as an Area Editor of the IEEE Transactions on Wireless Communications in resource management and multiple access from 2009 to 2011, an Editor of the IEEE Transactions on Mobile Computing from 2007 to 2012, and an Editor of the IEEE Journal on Selected Areas in Communications-Cognitive Radio Series from 2011 to 2014. He was the Distinguished Lecturer of the IEEE Communications Society from 2012 to 2015. He is currently a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is also a Registered Professional Engineer in the province of Manitoba, Canada.