Energy-Efficient Transmission in Heterogeneous Wireless Networks: A Delay-Aware Approach

Yuzhou Li, Yan Shi, *Member, IEEE*, Min Sheng, *Member, IEEE*, Cheng-Xiang Wang, *Senior Member, IEEE*, Jiandong Li, *Senior Member, IEEE*, Xijun Wang, *Member, IEEE*, and Yan Zhang, *Member, IEEE*

***Abstract*—In this paper, we investigate the delay-aware energy-efficient transmission problem in dynamic heterogeneous wireless networks (HWNs) with time-variant channel conditions, random traffic loads, and user mobility. By jointly considering subcarrier assignment, power allocation, and time fraction determination, we formulate it as a stochastic optimization problem to maximize the system energy efficiency (EE) and to ensure network stability. By leveraging the fractional programming theory and the Lyapunov optimization technique, we first propose a general algorithm framework, referred to as the eTrans, to solve the formulation. Further, we exploit the special structure of the subproblem embedded in the eTrans to develop the extremely simple and low complexity but optimal algorithms for subcarrier assignment, power allocation, and time fraction determination. In particular, all of them have closed-form solutions, and no iteration is required, which paves the way for employing the eTrans to practical applications. The theoretical analysis and simulation results exhibit that eTrans can flexibly strike a balance between EE and average delay by simply tuning an introduced control parameter.**

***Index Terms*—Delay, energy efficiency (EE), heterogeneous wireless networks (HWNs), power allocation, subcarrier assignment, time fraction determination.**

I. Introduction

GREEN radio, which aims for energy-efficient operation of wireless communication systems to combat ever-increasing energy consumption and carbon footprints, has achieved the worldwide recognition [1]–[3]. Meanwhile, heterogeneous wireless networks (HWNs), consisting of a diverse set of radio access networks (RANs) such as the Third-Generation Partnership Project Long-Term Evolution (3GPP LTE), Worldwide Interoperability for Microwave Access (WiMAX), and Wi-Fi, are becoming potential solutions against exponential traffic increase and ubiquitous wireless communications [4]–[6]. As a result, an important question that arises is how the system controller adaptively allocates limited wireless resource across all RANs to mobile terminals (MTs) to achieve ubiquitously energy-efficient transmission in such heterogeneous wireless environments.

There have been extensive studies on resource allocation in HWNs from different perspectives. Utility-based or throughput-based resource allocation was investigated in [7]–[9]. Bandwidth and power were jointly allocated in [10] to maximize the total system capacity by a formulated optimization problem. An analytical model to achieve the system sum-rate maximization subject to the proportional user rate constraint was presented in [11] to deal with the radio resource management. By determining power allocation and packet scheduling, in [12], an energy- and content-aware framework for multihoming video transmission to minimize video quality distortion was developed. As a common feature, in [7]–[12], network resource allocation schemes based on the stationary channel conditions and infinite backlog assumptions were devised. However, practical wireless networks operate in the presence of time-varying channel conditions and stochastic traffic arrivals. Moreover, delay performance was also neglected in [7]–[12], but it is an important metric to evaluate the quality of service (QoS) of traffic. Because of these, in [13] and [14], resource allocation algorithms that minimize power consumption (PC) in dynamic scenarios, with the average delay taken into account, were presented.

Unfortunately, all the aforementioned metrics, i.e., utility [7], [8], throughput [9], [10], spectral efficiency (SE) [11], distortion [12], and PC [13], [14], fail to measure energy efficiency (EE) [1]. As a specialized metric, EE is newly proposed to evaluate how efficiently the energy is consumed for energy-efficient (i.e., green) communications [1]–[3]. Some recent works have studied EE optimization problems [15]–[20]. In particular, an analytical framework was developed in [15] to evaluate the throughput and EE performance in downlink small-cell networks with multi-antenna base stations (BSs). A Markov chain was built in [16] to conduct performance analysis on SE and EE for two-tier femtocell networks with partially open channels. In [17]–[20], energy-efficient transmission problems were explored by jointly determining subcarrier assignment and power allocation. However, as in [7]–[12], in [15]–[20], problems were formulated subject to static models and did not consider mobility and delay as well. Moreover, in [7]–[19],

Manuscript received April 9, 2014; revised May 6, 2015; accepted July 23, 2015. Date of publication August 25, 2015; date of current version September 15, 2016. This work was supported by the National Natural Science Foundation of China under Grant 61231008, Grant 61172079, Grant 61201141, Grant 61301176, and Grant 91338114; by the 863 project under Grant 2014AA01A701; by the 111 Project under Grant B08038; by the Basic Research Fund in Xidian University under Grant 7214497201; and by the European Commission through the QUICK Project, International Research Staff Exchange Scheme under Grant FP7-PEOPLE-2013-IRSES. The review of this paper was coordinated by Dr. T. Taleb.

Y. Li, Y. Shi, M. Sheng, J. Li, X. Wang, and Y. Zhang are with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China (e-mail: yuzhou_li@stu.xidian.edu.cn; yshi@xidian.edu.cn; msheng@mail.xidian.edu.cn; jdli@mail.xidian.edu.cn; xijunwang@xidian.edu.cn; yanzhang@xidian.edu.cn).

C.-X. Wang is with the Institute of Sensors, Signals and Systems, School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, U.K., and also with the School of Information Science and Engineering, Shandong University, Jinan 250100, China (e-mail: cheng-xiang.wang@hw.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2015.2472578

subcarrier assignment, power allocation, or time fraction determination, or all of them, were not incorporated into the formulations.

To this end, in this paper, we address the delay-aware energy-efficient resource-allocation problem in HWNs by jointly considering time-variant channel conditions, random traffic loads, and user mobility. We formulate it as a stochastic optimization problem, which maximizes the system EE subject to the stability, subcarrier assignment, power allocation, and time fraction determination constraints. Using the fractional programming theory and the Lyapunov optimization technique, we first propose a general algorithm framework, referred to as the eTrans, to solve the formulation. In the eTrans, it is required that an optimization problem is solved with subcarrier assignment, power allocation, and time fraction being variables. We then develop optimal algorithms with low complexity to tackle it.

The main contributions of this paper are threefold.

- We combine multi-dimensional resource allocation and network stability to formulate the energy-efficient transmission problem in realistic HWNs, where time-variant channel conditions, random traffic loads, and user mobility are all taken into account.
- The devised eTrans provides a simple method to flexibly control EE–delay performance just by adjusting an introduced control parameter. In addition, the proposed method can be easily applied to other stochastic optimization problems with ratio-form objectives.
- By exploiting the special structure of the subproblem in the eTrans, we develop the extremely simple and low-complexity but optimal algorithms for subcarrier assignment, power allocation, and time fraction determination. In particular, all of them have closed-form solutions, and no iteration and optimization tools are required, which is of great importance for practical applications.

The remainder of this paper is organized as follows. In Section II, we introduce system scenarios and resource allocation models. Section III formulates the concerned problem and develops a general algorithm framework to solve the formulation. The optimal algorithms for subcarrier assignment, power allocation, and time fraction determination are devised in Section IV. We analyze the performance of the proposed algorithm in Section V. Simulation results are presented in Section VI. Finally, we conclude this paper in Section VII.

II. System Scenarios and Resource Allocation

In this section, after describing the concerned system scenarios, we elaborate the resource allocation in HWNs. We then present the definitions of network stability and EE, which will be used in Section III to formulate our considered problem.

A. Description of the System Scenario

The concerned HWNs, as shown in Fig. 1, are composed of the cellular network (CN) and Wi-Fi,[1] where users move. It is

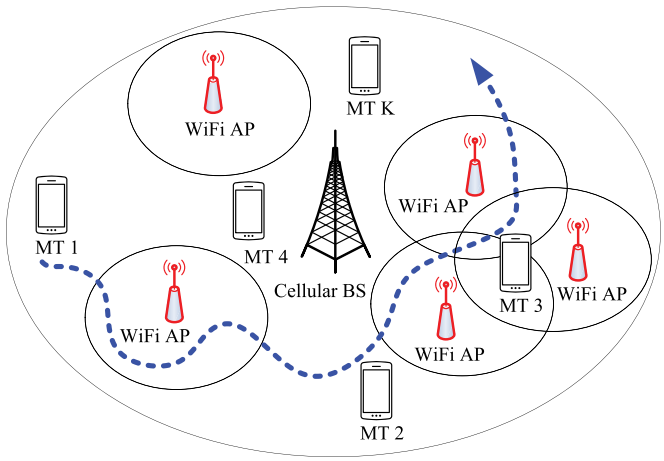


Fig. 1. Illustration scenario of HWNs.

one of the most popular integration scenarios in HWNs [4], [7], [13], [20]–[22] and is advocated by Qualcomm as an efficient network architecture to solve the future 1000× data challenge [6]. In this paper, we consider orthogonal frequency-division multiple access (OFDMA)-based cellular systems, e.g., the 3GPP LTE and WiMAX, and assume that the BS and each access point (AP) work on nonoverlapping channels; thus, there is no interference among the BS and APs [20], [23]. In comparison, the coverage of the CN is much larger, whereas Wi-Fi is more bandwidth efficient and energy efficient. Wi-Fi APs are usually deployed in the coverage of the CN BS. As a result, energy-expensive CN is always available, whereas Wi-Fi may not be available during the process of communication and movement of users. This indicates that the RANs that are able to serve MTs are time-variant, as shown by MT 1 in Fig. 1.

There are K active users in the system and N subcarriers in the CN. We denote the total bandwidth by B_{tot} and thus each subcarrier has a bandwidth of $W = B_{\text{tot}}/N$. Assuming that the network operates in slotted time with slots normalized to integral units; thus, slot t refers to the interval $[t, t+1)$, $t \in \{0, 1, 2, \ldots\}$. We denote the set of the available Wi-Fi APs for MT k by $\mathcal{N}_k(t)$ at time slot t. From Fig. 1, $\mathcal{N}_k(t)$ possibly includes no or one or several Wi-Fi APs. Furthermore, data arrives randomly at every slot and are queued separately for transmission to each receiver. Let $Q_k(t)$ and $A_k(t)$ denote the queue length and the new traffic arrival amount of MT k at slot t, respectively. Moreover, $\boldsymbol{A}(t) = (A_k(t))$ is assumed to be independent and identically distributed (i.i.d.) over time slots with $\mathbb{E}\{\boldsymbol{A}(t)\} = \boldsymbol{\lambda}$. In addition, $\boldsymbol{G}^{\text{BS}}(t) = (g_{k,n}(t))$ and $\boldsymbol{G}^{\text{AP}}(t) = (g_{k,m}(t))$ $(m \in \mathcal{N}_k(t))$ are channel gains, where $g_{k,n}(t)$ and $g_{k,m}(t)$ denote the channel gains from the BS to MT k on subcarrier n and from AP m to MT k at slot t, respectively. For simplicity, we assume that $\boldsymbol{G}(t) = [\boldsymbol{G}^{\text{BS}}(t), \boldsymbol{G}^{\text{AP}}(t)]$ takes values in a finite (but arbitrarily large) state space $\mathscr{G}$ with probabilities $\pi_{\boldsymbol{G}} = \pi_{\boldsymbol{G}_i}$, where $\pi_{\boldsymbol{G}_i} \triangleq \Pr[\boldsymbol{G}(t) = \boldsymbol{G}_i]$. We also assume that $\boldsymbol{G}(t)$ keeps constant for the duration of a time slot but potentially changes on slot boundaries.

B. Resource Allocation at the BS

Let $\boldsymbol{P}(t) = (P_{k,n}(t))$, where $P_{k,n}(t)$ denotes the transmit power from the BS to MT k on subcarrier n. We temporarily

[1]It is straightforward to extend the proposed model and algorithms to more types of RANs from the latter analysis.

assume that each subcarrier may be shared by multiple MTs in the time-division manner and denote by $\boldsymbol{\rho}(t) = (\rho_{k,n}(t)) \geq 0$ the time-sharing factor of MT k on subcarrier n. Here, it is worthwhile to note that, in Section IV, we will find that each subcarrier is allocated to at most one MT at optimality (cf. Theorem 4 and Remark 5). In other words, subcarriers are assigned exclusively among MTs, which is consistent with the standard assumption in OFDMA networks [24]–[28].

First, the transmit rate of MT k on subcarrier n (in unit of bit/s), $r_{k,n}(t)$, is given by [24]–[28]

$$r_{k,n}(t) = \rho_{k,n}(t) W \log_2 \left(1 + \frac{P_{k,n}(t)\,|h_{k,n}(t)|^2}{\rho_{k,n}(t) N_0 W}\right) = \rho_{k,n}(t) W \log_2 \left(1 + \frac{P_{k,n}(t) g_{k,n}(t)}{\rho_{k,n}(t)}\right) \quad (1)$$

where $g_{k,n}(t) = |h_{k,n}(t)|^2/(N_0 W)$, N_0 is the single-sided spectral density of additive white Gaussian noise, and $h_{k,n}(t)$ is the frequency response of link k on subcarrier n, which is assumed accurately known at the transmitter [18], [20], [24]–[28]. In addition, we set $r_{k,n}(t) = 0$ when $\rho_{k,n}(t) = 0$ in (1).

Accordingly, the transmit rate to MT k from the CN over all the subcarriers is equal to

$$R_k^{\mathrm{BS}}(t) = R_k^{\mathrm{BS}}\left(\boldsymbol{\rho}(t), \boldsymbol{P}(t)\right) = \sum_{n=1}^{N} r_{k,n}(t) \qquad \forall\, k. \quad (2)$$

In addition, the sum PC of the BS is

$$\mathrm{PC}_{\mathrm{tot}}^{\mathrm{BS}}(t) = \mathrm{PC}_{\mathrm{tot}}^{\mathrm{BS}}\left(\boldsymbol{\rho}(t), \boldsymbol{P}(t)\right) = \xi \sum_{k=1}^{K} \sum_{n=1}^{N} P_{k,n}(t) \quad (3)$$

where ξ is a constant that accounts for the inefficiency of the power amplifiers at the BS [26], [27], [29], [30]. Note that, in this paper, we do not incorporate the static power at the BS, which is consumed by the baseband signal processing and additional circuit blocks, such as analog-to-digital conversion, modulation, channel coding, and signal detection, into (3) as in [26] and [27], as we assume that the BS is always in the active state.

C. Resource Allocation at the APs

Regarding Wi-Fi, for the ease of analysis, we adopt the early backoff announcement (EBA), which is an improved version of distributed coordination function (DCF) with a reservation-based medium access control (MAC) protocol [31]. In the EBA, the collision among MTs can be completely avoided by announcing the future backoff information in the MAC header. As a result, we may regard that Wi-Fi operates in the time-division multiple access (TDMA) manner with each MT occupying the whole bandwidth in its allocated time fraction [20].

Let $P_{\mathrm{out}}^{\mathrm{AP}}$ denote the output power radiated at the transmit antenna of the AP, which is a constant and the same for all associated MTs. The transmit rate from AP m to MT k is given by $r_{k,m}(t)$, which is determined by the instantaneous signal-to-noise ratio (SNR) between them. Specifically

$$r_{k,m}(t) = f\left(\frac{P_{\mathrm{out}}^{\mathrm{AP}} g_{k,m}(t)}{N_0{}'}\right) \quad (4)$$

where $g_{k,m}(t)$ indicates the channel gain from AP m to MT k, $N_0{}'$ is the noise variance of the Wi-Fi channel, and $f(\cdot)$ decides the achievable data rate based on the SNR threshold [32].

As APs serve MTs in the TDMA manner, we denote by $\boldsymbol{X}(t) = (x_{k,m}(t))$ the time fraction of MT k served by AP m. We in this paper assume that each MT communicates at most with one AP that provides the highest data rate among all APs, then the transmit rate to MT k obtained from APs is[2]

$$R_k^{\mathrm{AP}}(t) = R_k^{\mathrm{AP}}\left(\boldsymbol{X}(t)\right) = \sum_{m=1}^{M} x_{k,m}(t) r_{k,m}(t). \quad (5)$$

In the following, we present the PC model for APs. An AP consumes constant power $P_{\mathrm{idle}}^{\mathrm{AP}}$ in the idle state and $P_{\mathrm{tx}}^{\mathrm{AP}}$ in the transmit state.[3] Thus, the total PC of AP m can be modeled as [23]

$$\mathrm{PC}_m^{\mathrm{AP}}(t) = \sum_{k=1}^{K} x_{k,m}(t) P_{\mathrm{tx}}^{\mathrm{AP}} + \left(1 - \sum_{k=1}^{K} x_{k,m}(t)\right) P_{\mathrm{idle}}^{\mathrm{AP}}. \quad (6)$$

In (6), the first and second terms denote the power consumed at AP m in the active and idle states, respectively.

D. Definitions of EE and Network Stability

From (2) and (5), the sum transmit rate to MT k provided by the system is

$$R_k(t) = R_k^{\mathrm{BS}}(t) + R_k^{\mathrm{AP}}(t). \quad (7)$$

In addition, the total PC and transmit rate of the overall network are, respectively, given by

$$\mathrm{PC}_{\mathrm{tot}}(t) = \mathrm{PC}_{\mathrm{tot}}^{\mathrm{BS}}(t) + \sum_{m=1}^{M} \mathrm{PC}_m^{\mathrm{AP}}(t) \quad (8)$$

$$R_{\mathrm{tot}}(t) = \sum_{k=1}^{K} R_k(t). \quad (9)$$

Furthermore, we define the time averaged PC and transmit rate of the system, respectively, as

$$\overline{\mathrm{PC}}_{\mathrm{tot}} = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}(\tau)\right\} \quad (10)$$

$$\overline{R}_{\mathrm{tot}} = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{R_{\mathrm{tot}}(\tau)\right\}. \quad (11)$$

[2]We use the sum operation in (5) to keep the consistence with the notations in OFDMA networks but it is worthwhile to point out that there is at most one term in the sum expression due to our association assumption between APs and MTs.

[3]Note that $P_{\mathrm{tx}}^{\mathrm{AP}}$ denotes the total power consumed in the transmit mode, including the output power $P_{\mathrm{out}}^{\mathrm{AP}}$.

Here, we assume temporarily that the limits are well defined.

To quantitatively model the impacts of stochastic traffic arrivals, time-varying channel conditions, and unpredicted available RANs on the resource allocation policy $\boldsymbol{\rho}(t)$, $\boldsymbol{P}(t)$, and $\boldsymbol{X}(t)$, and further on EE and delay, we give the precise definitions of EE and network stability.

Definition 1: EE η_{EE} of networks is defined as the ratio of the long-term delivered amount of data to the corresponding energy consumption in units of bit/Joule [33], and given by

$$\eta_{\mathrm{EE}} = \lim_{t\to\infty} \frac{\sum_{\tau=0}^{t-1} \mathbb{E}\left\{R_{\mathrm{tot}}(\tau)\right\}}{\sum_{\tau=0}^{t-1} \mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}(\tau)\right\}} = \frac{\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{R_{\mathrm{tot}}(\tau)\right\}}{\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}(\tau)\right\}} = \frac{\overline{R}_{\mathrm{tot}}}{\overline{\mathrm{PC}}_{\mathrm{tot}}}. \tag{12}$$

Let $\boldsymbol{\rho}=\{\boldsymbol{\rho}(0),\boldsymbol{\rho}(1),\boldsymbol{\rho}(2),\ldots\}$, $\boldsymbol{P}=\{\boldsymbol{P}(0),\boldsymbol{P}(1),\boldsymbol{P}(2),\ldots\}$, and $\boldsymbol{X}=\{\boldsymbol{X}(0),\boldsymbol{X}(1),\boldsymbol{X}(2),\ldots\}$. Indeed, both $\overline{\mathrm{PC}}_{\mathrm{tot}}$ and $\overline{R}_{\mathrm{tot}}$ are the functions of $\boldsymbol{\rho}$, $\boldsymbol{P}$, and $\boldsymbol{X}$ and thus can be expressed as $\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})$ and $\overline{R}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})$, respectively.

Remark 1: The EE definition (12) is quite different from those presented in the existing works [18], [20], [26]–[28] (and the references therein). We refer the readers to [33] on the detailed analysis for their connections and differences.

As stated earlier, the traffic arrival and departure rates of MT k are $A_k(t)$ and $R_k(t)$, respectively. Thus, the data queue $Q(t)$ evolves according to [34], [35]

$$Q_k(t+1) = \max\left[Q_k(t) - R_k(t), 0\right] + A_k(t). \tag{13}$$

To achieve energy-efficient transmission, an intuitive strategy is to transmit data only when the channel conditions from the BS or APs to MTs are good enough. However, if the controller greedily defers data transmission for the aggressive consideration on EE, the data queue length is likely to increase unboundedly, thus resulting in a large intolerant delay. As a result, the controller falls into a dilemma to decide whether to allocate resource or not to MTs for transmission in each slot. This, in essence, is an EE–delay tradeoff issue. We thus introduce queue stability to flexibly balance or control such a tradeoff.

Definition 2: A discrete time queue process $Q(t)$ is strongly stable [34] if

$$\overline{Q} = \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{|Q(t)|\right\} < \infty. \tag{14}$$

Definition 3: A network of queues is stable if all the individual queues are stable [34].

III. Problem Formulation and Algorithm Framework

In this section, we first formulate a stochastic optimization programming to investigate the energy-efficient transmission problem, which maximizes EE subject to resource allocation constraints and a bounded delay requirement. We then adopt the fractional programming theory to equivalently reform the proposed formulation. Finally, we devise a general algorithm framework, referred to as the eTrans, to solve our problem by using the Lyapunov optimization technique.

A. Problem Formulation

In this paper, with delay taken into account, we joint subcarrier assignment, power allocation, and time fraction determination to investigate the energy-efficient transmission problem in HWNs. We formulate it as the following stochastic optimization programming:

$$\begin{aligned}
\max \quad & \eta_{\mathrm{EE}} = \frac{\overline{R}_{\mathrm{tot}}}{\overline{\mathrm{PC}}_{\mathrm{tot}}} = \frac{\overline{R}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})}{\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})} \\
\text{s.t.} \quad & \mathrm{C1}: \overline{Q}_k < \infty \qquad \forall\, k \\
& \mathrm{C2}: \sum_{k=1}^{K} \rho_{k,n}(t) \le 1 \qquad \forall\, n,t \\
& \mathrm{C3}: \sum_{k=1}^{K} x_{k,m}(t) \le 1 \qquad \forall\, m,t \\
& \mathrm{C4}: 0 \le \rho_{k,n}(t) \le 1 \qquad \forall\, k,n,t \\
& \mathrm{C5}: 0 \le x_{k,m}(t) \le 1 \qquad \forall\, k,m,t \\
& \mathrm{C6}: P_{k,n}(t) \ge 0 \qquad \forall\, k,n,t.
\end{aligned} \tag{15}$$

In (15), C1 is the queue stability constraint to guarantee all arrived data leaving the buffer within a finite time. C2 denotes the time sharing constraints on each subcarrier among MTs at the BS. C3 represents the time fraction constraints among MTs at APs. Note that we depict the average delay by the queue length (i.e., C1). This is because the average delay is proportional to the average queue length for a given traffic arrival rate from the Little's Theorem [36] (i.e., average delay = average queue length/traffic arrival rate).

B. Problem Transformation

We cannot directly employ the classical drift-plus-penalty algorithm [34], which is developed to solve stochastic optimization problems by the Lyapunov optimization technique, to tackle (15) since its objective is the ratio between two time average quantities rather than a time average of a quantity. However, (15) falls into the kind of nonlinear fractional programming [37], which can be exploited to transform the ratio to a time average of a quantity.

For the notational simplicity, we define $\boldsymbol{\mathcal{R}}$ as the set of feasible solutions of (15). Without loss of generality, we have $\overline{R}_{\mathrm{tot}} > 0$ and $\overline{\mathrm{PC}}_{\mathrm{tot}} > 0$. Denote the optimal value of (15) by $\eta_{\mathrm{EE}}^{\mathrm{opt}}$; then

$$\eta_{\mathrm{EE}}^{\mathrm{opt}} = \frac{\overline{R}_{\mathrm{tot}}(\boldsymbol{\rho}^*,\boldsymbol{P}^*,\boldsymbol{X}^*)}{\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho}^*,\boldsymbol{P}^*,\boldsymbol{X}^*)} = \max_{\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X}\in\boldsymbol{\mathcal{R}}} \frac{\overline{R}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})}{\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})} \tag{16}$$

where $\boldsymbol{\rho}^*$, $\boldsymbol{P}^*$, and $\boldsymbol{X}^*$ refer to the optimal resource allocation policy.

We first present the following theorem to reformulate (15). Its proof uses a standard result in the nonlinear fractional programming theory [37], we thus omit it for brevity.

Theorem 1: The resource allocation policy $\boldsymbol{\rho}^*$, $\boldsymbol{P}^*$, and $\boldsymbol{X}^*$ achieves the optimal EE $\eta_{\mathrm{EE}}^{\mathrm{opt}}$ if and only if

$$\begin{aligned}&\max_{\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X}\in\mathcal{R}} \overline{R}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X}) - \eta_{\mathrm{EE}}^{\mathrm{opt}}\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})\\ &\quad= \overline{R}_{\mathrm{tot}}(\boldsymbol{\rho}^*,\boldsymbol{P}^*,\boldsymbol{X}^*) - \eta_{\mathrm{EE}}^{\mathrm{opt}}\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho}^*,\boldsymbol{P}^*,\boldsymbol{X}^*)\\ &\quad= 0. \end{aligned} \tag{17}$$

From Theorem 1, given $\eta_{\mathrm{EE}}^{\mathrm{opt}}$, we can equivalently transform (15) to

$$\begin{aligned}\max \quad & \overline{R}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X}) - \eta_{\mathrm{EE}}^{\mathrm{opt}}\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})\\ \text{s.t.} \quad & \mathrm{C1},\mathrm{C2},\mathrm{C3},\mathrm{C4},\mathrm{C5},\mathrm{C6}.\end{aligned} \tag{18}$$

Although the objective of (18) can be seen as the time average of a quantity (i.e., $R_{\mathrm{tot}}(t) - \eta_{\mathrm{EE}}^{\mathrm{opt}}\mathrm{PC}_{\mathrm{tot}}(t)$), it is still challenging to solve (18) as $\eta_{\mathrm{EE}}^{\mathrm{opt}}$ is usually unknown in advance.

Further, we introduce a new parameter $\eta_{\mathrm{EE}}(t)$ with $\eta_{\mathrm{EE}}(0) = 0$ and define it as

$$\eta_{\mathrm{EE}}(t) = \frac{\sum_{\tau=0}^{t-1}\mathbb{E}\left\{R_{\mathrm{tot}}(\tau)\right\}}{\sum_{\tau=0}^{t-1}\mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}(\tau)\right\}}, \quad t = 1,2,3,\ldots. \tag{19}$$

Then, replacing $\eta_{\mathrm{EE}}^{\mathrm{opt}}$ by $\eta_{\mathrm{EE}}(t)$, we recast (18) as

$$\begin{aligned}\max \quad & \overline{R}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X}) - \eta_{\mathrm{EE}}(t)\overline{\mathrm{PC}}_{\mathrm{tot}}(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{X})\\ \text{s.t.} \quad & \mathrm{C1},\mathrm{C2},\mathrm{C3},\mathrm{C4},\mathrm{C5},\mathrm{C6}.\end{aligned} \tag{20}$$

Depending on the past resource allocation decisions, $\eta_{\mathrm{EE}}(t)$, unlike $\eta_{\mathrm{EE}}^{\mathrm{opt}}$, becomes a known parameter. This kind of transformation techniques has been widely used to solve stochastic optimization problems with ratio objectives in renewal systems and has been shown to be extremely effective [34], [38].

C. Delay-Aware Energy-Efficient Transmission Algorithm (eTrans)

Hereto, the objective of (20) can be seen as the time average of $R_{\mathrm{tot}}(t) - \eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t)$; thus, we can exploit the drift-plus-penalty algorithm to solve (20), i.e., (15). To this end, we first introduce some necessary but practical boundedness assumptions to derive the drift-plus-penalty expression of (20).

Under any channel condition $\boldsymbol{G}(t)$, any queue state $\boldsymbol{Q}(t) = (Q_k(t))$, and the corresponding resource allocation policy $\boldsymbol{\rho}(t),\boldsymbol{P}(t),\boldsymbol{X}(t) \in A_{\boldsymbol{W}(t)}$ (where $\boldsymbol{W}(t) = [\boldsymbol{G}(t),\boldsymbol{Q}(t)]$, and $A_{\boldsymbol{W}(t)}$ represents the set of all resource allocation options for a given $\boldsymbol{W}(t)$), we assume

$$\mathrm{PC}_{\min} \le \mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}(t)\right\} \le \mathrm{PC}_{\max} \tag{21}$$

$$R_{\min} \le \mathbb{E}\left\{R_{\mathrm{tot}}(t)\right\} \le R_{\max} \tag{22}$$

where $\mathrm{PC}_{\min}$, $\mathrm{PC}_{\max}$, $R_{\min}$, and $R_{\max}$ are some finite constants. Moreover, assume that the traffic arrival amount $\boldsymbol{A}(t) = (A_k(t))$ has a bounded second moment, i.e.,

$$\mathbb{E}\left\{\boldsymbol{A}(t)^2\right\} \le \mathscr{A} \tag{23}$$

for some finite constant $\mathscr{A}$.

It is worthwhile to note that the assumptions (21)–(23) are reasonable due to bounded arrival/departure rates and power allocation in realistic systems. Moreover, these lower and upper bounds will not be used when developing algorithms.

In what follows, we introduce the concepts of the Lyapunov function and Lyapunov drift, which will also be used in Lemma 1 to derive the drift-plus-penalty expression. From [34], the Lyapunov function $L(\boldsymbol{Q}(t))$ and the one-slot conditional Lyapunov drift $\Delta(\boldsymbol{Q}(t))$ are, respectively, defined as

$$L(\boldsymbol{Q}(t)) \triangleq \frac{1}{2}\sum_{k=1}^{K} Q_k(t)^2 \tag{24}$$

$$\Delta\left(\boldsymbol{Q}(t)\right) \triangleq \mathbb{E}\left\{L\left(\boldsymbol{Q}(t+1)\right) - L\left(\boldsymbol{Q}(t)\right)|\boldsymbol{Q}(t)\right\} \tag{25}$$

and the drift-plus-penalty expression of (20) is given by

$$\Delta\left(\boldsymbol{Q}(t)\right) + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)|\boldsymbol{Q}(t)\right\}. \tag{26}$$

The following Lemma offers an upper bound of the drift-plus-penalty expression. The proof of this lemma uses a standard method in the stochastic optimization theory [33], [34]. For example, without considering the virtual power queue introduced in [33], the method adopted in [33, Lemma 2, p. 6] can be easily applied to prove this lemma. More specifically, we complete its proof by removing (68) and (69) in [33, Appendix C, pp. 9–10].

Lemma 1: For any subcarrier assignment, power allocation, and time fraction determination algorithm, all control parameters $V \ge 0$, and all possible values of $\boldsymbol{Q}(t)$, the drift-plus-penalty expression of (20) is upper bounded by

$$\begin{aligned}&\Delta\left(\boldsymbol{Q}(t)\right) + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)|\boldsymbol{Q}(t)\right\}\\ &\quad\le B + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)|\boldsymbol{Q}(t)\right\}\\ &\qquad+ \sum_{k=1}^{K} Q_k(t)\mathbb{E}\left\{A_k(t) - R_k(t)|\boldsymbol{Q}(t)\right\}\end{aligned} \tag{27}$$

where B is a positive constant [according to the boundedness assumptions (21)–(23)] that, for all t, satisfies

$$B \ge \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left\{A_k(t)^2 + R_k(t)^2|\boldsymbol{Q}(t)\right\}. \tag{28}$$

From the stochastic optimization theory in [34], it is required to minimize the upper bound of its drift-plus-penalty expression subject to the same constraints, except the stability one to solve a stochastic optimization problem. Back to our case, we need to minimize the right-hand side of (27), i.e., the upper bound of the drift-plus-penalty of (20), to solve (20) subject to C2–C6, as C1 is a stability constraint. Algorithm 1, which is referred

to as the eTrans in this paper,[4] provides the detailed procedure to deal with (20) [i.e., (15)]. Regarding its performance, the eTrans can achieve an asymptotically optimal EE that arbitrarily approaches the theoretical optimum of (15) at the cost of the average delay. However, to avoid tedious derivations, we leave the detailed performance analysis in Section V.

Algorithm 1 Delay-aware energy-efficient transmission algorithm (eTrans).

1: At the beginning of each slot t, observe the current queue state $\boldsymbol{Q}(t)$, obtain the channel condition $\boldsymbol{G}(t)$, and get the available APs $\mathcal{N}_k(t)$ for all k.

2: Make the resource allocation policy $\boldsymbol{\rho}(t), \boldsymbol{P}(t)$, and $\boldsymbol{X}(t)$ according to the following optimization problem:

$$\begin{aligned}\min \quad & V[\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)] + \sum_{k=1}^{K} Q_k(t)[A_k(t) - R_k(t)] \\ \text{s.t.} \quad & \mathrm{C2}: \sum_{k=1}^{K} \rho_{k,n}(t) \le 1 \qquad \forall\, n,t \\ & \mathrm{C3}: \sum_{k=1}^{K} x_{k,m}(t) \le 1 \qquad \forall\, m,t \\ & \mathrm{C4}: 0 \le \rho_{k,n}(t) \le 1 \qquad \forall\, k,n,t \\ & \mathrm{C5}: 0 \le x_{k,m}(t) \le 1 \qquad \forall\, k,m,t \\ & \mathrm{C6}: P_{k,n}(t) \ge 0 \qquad \forall\, k,n,t. \end{aligned} \tag{29}$$

3: Update $\boldsymbol{Q}(t)$ and $\eta_{\mathrm{EE}}(t)$ according to (13) and (19), respectively.

Remark 2: Recall that $\eta_{\mathrm{EE}}(t)$ is a constant and $\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)$ can be interpreted as a metric to measure the system EE at slot t. In addition, V is an introduced positive control parameter. An intuitional explanation to the physical meaning of (29) is that it minimizes the weighted sum (by V) between the system EE and the weighted (by queue length, i.e., delay) transmit rate subject to resource allocation constraints. In other words, (29) couples two targets, i.e., EE and delay, and balances them by adjusting V. Particularly, the eTrans requires no prior knowledge of traffic arrival rate $\boldsymbol{\lambda}$, channel statistics $\pi_{\boldsymbol{G}}$, and available RANs in the future.

[4]Note that the conditional expectation is removed from (27) to (29) due to the fact that minimizing $f(t)$ ensures that $\mathbb{E}\{f(t)|\boldsymbol{Q}(t)\}$ is minimized from the principle of opportunistically minimizing an expectation [34].

IV. Optimal Resource Allocation Algorithms

In the last section, we have developed the eTrans to solve our proposed formulation (15). As shown, it is required to devise efficient algorithms to tackle (29) for practical applications. In this section, we design extremely simple and low-complexity but optimal algorithms for subcarrier assignment, power allocation, and time fraction allocation, all of which have closed-form solutions.

By substituting (7)–(9) into (29), we obtain (30), shown at the bottom of the page. Furthermore, we can decompose (30) into two subproblems: one for time fraction allocation and the other for subcarrier assignment and power allocation. This is because there are no coupling objective functions and constraints between them.

A. Optimal Time Fraction Allocation

The time fraction allocation problem is given as follows:

$$\begin{aligned}\min \quad & \sum_{k=1}^{K}\sum_{m=1}^{M} \big[V\eta_{\mathrm{EE}}(t)P_{\mathrm{tx}}^{\mathrm{AP}} - V\eta_{\mathrm{EE}}(t)P_{\mathrm{idle}}^{\mathrm{AP}} \\ & \qquad - (V + Q_k(t))\, r_{k,m}(t)\big] x_{k,m}(t) \\ \text{s.t.} \quad & \mathrm{C3}: \sum_{k=1}^{K} x_{k,m}(t) \le 1 \qquad \forall\, m,t \\ & \mathrm{C5}: 0 \le x_{k,m}(t) \le 1 \qquad \forall\, k,m,t. \end{aligned} \tag{31}$$

$$\begin{aligned}\min \quad & V\eta_{\mathrm{EE}}(t)P_{\mathrm{static}}^{\mathrm{BS}} + MV\eta_{\mathrm{EE}}(t)P_{\mathrm{idle}}^{\mathrm{AP}} + \sum_{k=1}^{K} Q_k(t)A_k(t) + \sum_{k=1}^{K}\sum_{n=1}^{N} [\xi V\eta_{\mathrm{EE}}(t)P_{k,n}(t) - (V + Q_k(t))\, r_{k,n}(t)] \\ & + \sum_{k=1}^{K}\sum_{m=1}^{M} \left[V\eta_{\mathrm{EE}}(t)P_{\mathrm{tx}}^{\mathrm{AP}} - V\eta_{\mathrm{EE}}(t)P_{\mathrm{idle}}^{\mathrm{AP}} - (V + Q_k(t))\, r_{k,m}(t)\right] x_{k,m}(t) \\ \text{s.t.} \quad & \mathrm{C2}: \sum_{k=1}^{K} \rho_{k,n}(t) \le 1 \qquad \forall\, n,t \\ & \mathrm{C3}: \sum_{k=1}^{K} x_{k,m}(t) \le 1 \qquad \forall\, m,t \\ & \mathrm{C4}: 0 \le \rho_{k,n}(t) \le 1 \qquad \forall\, k,n,t \\ & \mathrm{C5}: 0 \le x_{k,m}(t) \le 1 \qquad \forall\, k,m,t \\ & \mathrm{C6}: P_{k,n}(t) \ge 0 \qquad \forall\, k,n,t \end{aligned} \tag{30}$$

Recall that each MT is associated with at most one AP, which provides the highest transmit rate among all the APs. Thus, if MT k communicates with AP m at slot t, i.e., $r_{k,m}(t) \geq r_{k,j}(t)$ for $\forall\, j \neq m$, then $x_{k,j}(t) = 0$, whereas the value of $x_{k,m}(t)$ is determined by (31). If $x_{k,m}(t) = 0$, then MT k gets no service from any AP at this slot. We denote the set of MTs that are possibly[5] connected to AP m by $\Omega_m(t)$, i.e., $\Omega_m(t) = \{k | r_{k,m}(t) \geq r_{k,j}(t), \forall\, j \neq m\}$. Note that an AP can serve multiple MTs in a slot in the TDMA manner from C3.

For the notational simplicity, we let

$$\alpha_{k,m}(t) = V\eta_{\mathrm{EE}}(t)\left(P_{\mathrm{tx}}^{\mathrm{AP}} - P_{\mathrm{idle}}^{\mathrm{AP}}\right) - \left(V + Q_k(t)\right) r_{k,m}(t). \tag{32}$$

Then, regarding the optimal solution of (31), we present the following theorem.

Theorem 2: For any MT $k \in \Omega_m(t)$, $x_{k,j}(t) = 0$ for all $j \neq m$, and the value of $x_{k,m}(t)$ is determined by

$$x_{k,m}(t) = \begin{cases} 0, & \text{if } k \in \Omega_m^+(t) \\ 1, & \text{if } k \in \Omega_m^-(t) \text{ and } k = k^* \\ 0, & \text{if } k \in \Omega_m^-(t) \text{ and } k \neq k^* \end{cases} \tag{33}$$

where

$$\begin{aligned} \Omega_m^+(t) &= \{k | \alpha_{k,m} \geq 0, k \in \Omega_m(t)\} \\ \Omega_m^-(t) &= \{k | \alpha_{k,m} < 0, k \in \Omega_m(t)\} \\ k^* &= \arg \min_{i \in \Omega_m^-(t)} \alpha_{i,m}(t). \end{aligned} \tag{34}$$

Proof: Based on the given analysis, $x_{k,j}(t) = 0$ for $k \in \Omega_m(t)$ and $j \neq m$. To show the later claims in Theorem 2, we first equivalently recast (31) to

$$\begin{aligned} \min \quad & \sum_{m=1}^{M} \sum_{k \in \Omega_m} \alpha_{k,m}(t) x_{k,m}(t) \\ \text{s.t.} \quad & \mathrm{C3}: \sum_{k \in \Omega_m} x_{k,m}(t) \leq 1 \qquad \forall\, m, t \\ & \mathrm{C5}: 0 \leq x_{k,m}(t) \leq 1 \qquad \forall\, k, m, t. \end{aligned} \tag{35}$$

As shown in (35), the time fraction allocation among APs is independent. Thus, (35) can be decomposed into M subproblems, each of which (e.g., m) is given as follows:

$$\begin{aligned} \min \quad & \sum_{k \in \Omega_m} \alpha_{k,m}(t) x_{k,m}(t) \\ \text{s.t.} \quad & \mathrm{C3}: \sum_{k \in \Omega_m} x_{k,m}(t) \leq 1 \qquad \forall\, t \\ & \mathrm{C5}: 0 \leq x_{k,m}(t) \leq 1 \qquad \forall\, k \in \Omega_m, t. \end{aligned} \tag{36}$$

To minimize the objective function, it is obvious to set $x_{k,m}(t) = 0$ for MTs with $\alpha_{k,m}(t) \geq 0$. In other words, $x_{k,m}(t) = 0$ if $k \in \Omega_m^+(t)$. For MTs $k \in \Omega_m^-(t)$, e.g., k_1 and k_2, we have $x_{k_1,m}(t) \leq 1 - x_{k_2,m}(t)$ from C3. The objective of (36) is equal to

$$\begin{aligned} & \alpha_{k_1,m}(t) x_{k_1,m}(t) + \alpha_{k_2,m}(t) x_{k_2,m}(t) \\ & \geq \alpha_{k_1,m}(t)\left(1 - x_{k_2,m}(t)\right) + \alpha_{k_2,m}(t) x_{k_2,m}(t) \\ & \geq \left(\alpha_{k_2,m}(t) - \alpha_{k_1,m}(t)\right) x_{k_2,m}(t) + \alpha_{k_1,m}(t) \end{aligned} \tag{37}$$

where the first equality holds from the nonnegativity of $x_{k,m}(t)$ (see C5) and the negativity of $\alpha_{k,m}(t)$. To minimize $(\alpha_{k_2,m}(t) - \alpha_{k_1,m}(t)) x_{k_2,m}(t) + \alpha_{k_1,m}(t)$, we have $x_{k_2,m}(t) = 1$ and $x_{k_1,m}(t) = 0$ if $\alpha_{k_2,m}(t) < \alpha_{k_1,m}(t) < 0$. Otherwise, $x_{k_2,m}(t) = 0$ and $x_{k_1,m}(t) = 1$ if $\alpha_{k_1,m}(t) < \alpha_{k_2,m}(t) < 0$. Similar methods can be easily extended to the cases with more MTs in $\Omega_m^-(t)$. ■

Remark 3: The solution structure in (33) is completely in accordance with our intuitional understanding. 1) It should turn on AP m for a good EE or a low delay when $r_{k,m}(t)$ or $Q_k(t)$ becomes large. From (32), a large $r_{k,m}(t)$ or $Q_k(t)$ yields an increased second term in $\alpha_{k,m}(t)$ and a negative $\alpha_{k,m}(t)$, thus leading to $x_{k,m}(t) = 1$ [see (33)]. 2) AP m will always be turned on if $P_{\mathrm{tx}}^{\mathrm{AP}} = P_{\mathrm{tx}}^{\mathrm{AP}}$, as the on operation would not cause additional PC. It is observed that $x_{k,m}(t) = 1$ as $\alpha_{k,m}(t)$ is always negative when $P_{\mathrm{tx}}^{\mathrm{AP}} = P_{\mathrm{tx}}^{\mathrm{AP}}$. 3) It is more likely to turn on APs when $P_{\mathrm{tx}}^{\mathrm{AP}}$ approaches to $P_{\mathrm{idle}}^{\mathrm{AP}}$ and to turn off them when $P_{\mathrm{tx}}^{\mathrm{AP}}$ is larger than $P_{\mathrm{idle}}^{\mathrm{AP}}$.

B. Optimal Subcarrier Assignment and Power Allocation

Accordingly, the subcarrier assignment and power allocation problem is determined by

$$\begin{aligned} \min_{\boldsymbol{P}(t), \boldsymbol{\rho}(t)} \quad & \sum_{k=1}^{K} \sum_{n=1}^{N} \left[\xi V \eta_{\mathrm{EE}}(t) P_{k,n}(t) - \left(V + Q_k(t)\right) r_{k,n}(t)\right] \\ \text{s.t.} \quad & \mathrm{C2}: \sum_{k=1}^{K} \rho_{k,n}(t) \leq 1 \qquad \forall\, n, t \\ & \mathrm{C4}: 0 \leq \rho_{k,n}(t) \leq 1 \qquad \forall\, k, n, t \\ & \mathrm{C6}: P_{k,n}(t) \geq 0 \qquad \forall\, k, n, t. \end{aligned} \tag{38}$$

We present the following proposition, which is proved in the Appendix, to show the feature of (38).

Theorem 3: Problem (38) is jointly convex in $\boldsymbol{P}(t)$ and $\boldsymbol{\rho}(t)$.

Due to its convexity, we can adopt standard convex optimization techniques such as interior-point methods [39] to effectively and optimally solve (38) by off-the-shelf solvers, e.g., CVX [40]. However, we find that we can exploit the special structure of (38) to devise extremely simple closed-form solutions with substantially low complexity. For this aim, we first provide the following basic fact in the optimization theory.

Lemma 2 ([39, p. 133]): We always have

$$\inf_{x,y} f(x,y) = \inf_{x} \widetilde{f}(x) \tag{39}$$

where $\widetilde{f}(x) = \inf_y f(x,y)$. In other words, we can always minimize a function by first minimizing over some of the variables and then minimizing over the remaining ones.

From Lemma 2, we can solve (38) by first optimizing $\boldsymbol{P}(t)$ and then $\boldsymbol{\rho}(t)$. The detailed process is described as follows.

[5]Here, we use "possibly" because the MT (e.g., k) may get no service from the AP (e.g., m), even when it provides the highest rate, which depends on whether $x_{k,m}(t) = 0$ or not from (31).

Define the function $\widetilde{f}_0$ of $\boldsymbol{\rho}(t)$ by

$$\begin{aligned}\min_{\boldsymbol{P}(t)} \quad & \widetilde{f}_0\left(\boldsymbol{\rho}(t)\right) = \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\xi V \eta_{\mathrm{EE}}(t) P_{k,n}(t)\right. \\ & \qquad \left. - \left(V + Q_k(t)\right) r_{k,n}(t)\right] \\ \text{s.t.} \quad & \mathrm{C6}: P_{k,n}(t) \geq 0 \qquad \forall\, k, n, t. \end{aligned} \tag{40}$$

The problem (38) is then equivalent to

$$\begin{aligned}\min_{\boldsymbol{\rho}(t)} \quad & \widetilde{f}_0\left(\boldsymbol{\rho}(t)\right) \\ \text{s.t.} \quad & \mathrm{C2}: \sum_{k=1}^{K} \rho_{k,n}(t) \leq 1 \qquad \forall\, n, t \\ & \mathrm{C4}: 0 \leq \rho_{k,n}(t) \leq 1 \qquad \forall\, k, n, t. \end{aligned} \tag{41}$$

To obtain the optimal power allocation $P_{k,n}(t)$ from (40), we differentiate $\widetilde{f}_0(\boldsymbol{\rho}(t))$ with respect to $P_{k,n}(t)$ and set it to zero to yield

$$P_{k,n}(t) = \left[\frac{\left(V + Q_k(t)\right) W}{\xi V \eta_{\mathrm{EE}(t)} \ln 2} - \frac{1}{g_{k,n}(t)}\right]^{+} \rho_{k,n}(t) \quad \forall\, k, n, t. \tag{42}$$

Note that $[x]^{+} \triangleq \max[0, x]$.

Remark 4: As shown, the optimal power allocation (42) follows the standard water-filling approach. That is, the better the channel condition $g_{k,n}(t)$, the higher the transmit power $P_{k,n}(t)$. In addition, the water lever is determined by the queue length $Q_k(t)$ in the current slot, which perfectly agrees with our intuitional understanding. Specifically, it is necessary to pour more transmit power as $Q_k(t)$ increases because a larger transmit rate is required to keep the network stable.

By substituting the optimal power allocation (42) into $\widetilde{f}_0(\boldsymbol{\rho}(t))$, we equivalently recast (41) to

$$\begin{aligned}\min_{\boldsymbol{\rho}(t)} \quad & \widetilde{f}_0\left(\boldsymbol{\rho}(t)\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \varphi_{k,n}(t) \rho_{k,n}(t) \\ \text{s.t.} \quad & \mathrm{C2}: \sum_{k=1}^{K} \rho_{k,n}(t) \leq 1 \qquad \forall\, n, t \\ & \mathrm{C4}: 0 \leq \rho_{k,n}(t) \leq 1 \qquad \forall\, k, n, t \end{aligned} \tag{43}$$

where

$$\begin{aligned}\varphi_{k,n}(t) = & \left[\frac{\left(V + Q_k(t)\right) W}{\ln 2} - \frac{\xi V \eta_{\mathrm{EE}(t)}}{g_{k,n}(t)}\right]^{+} \\ & - \left(V + Q_k(t)\right) W \left[\log_2\left(\frac{\left(V + Q_k(t)\right) W g_{k,n}(t)}{\xi V \eta_{\mathrm{EE}(t)} \ln 2}\right)\right]^{+}. \end{aligned} \tag{44}$$

The following theorem specifies the optimal solution of (43), i.e., the optimal subcarrier assignment.

Theorem 4: For any given subcarrier n, the optimal subcarrier assignment $\rho_{k,n}(t)$ at slot t is given by

$$\rho_{k,n}(t) = \begin{cases} 0, & \text{if } \varphi_{k,n} \geq 0 \\ 1, & \text{if } \varphi_{k,n} < 0 \text{ and } k = \arg\min_i \varphi_{i,n}(t) \\ 0, & \text{if } \varphi_{k,n} < 0 \text{ and } k \neq \arg\min_i \varphi_{i,n}(t). \end{cases} \tag{45}$$

Proof: We can exploit the similar method adopted in Theorem 2 to prove Theorem 4, we thus omit the proof for brevity. ■

Remark 5: Although we permit that MTs can share subcarriers in the time-division manner in Section II-B, i.e., $\rho_{k,n}(t) \in [0, 1]$, the optimal scheme for subcarrier assignment from Theorem 4 is exclusive occupancy, i.e., each subcarrier will be assigned to at most one MT.

Jointly considering (33), (45), and (42), we develop Algorithm 2 to optimally determine time fractions, assign subcarriers, and allocation power, i.e., to optimally solve (29).

Algorithm 2 Low complexity but optimal time fraction determination, subcarrier assignment, and power allocation algorithm.

1: Determine time fraction $x_{k,n}(t)$ from (33).
2: Assign subcarrier $\rho_{k,n}(t)$ from (45).
3: Allocate transmit power $P_{k,n}(t)$ from (42).

V. Performance Analysis

In this section, we exploit the Lyapunov optimization technique to analyze the performance of the eTrans, where Algorithm 2 is called to optimally solve (29).

A. Preliminary

Assume that the time averages of the PC, rate, and EE converge. Specifically

$$\lim_{J\to\infty} \frac{1}{J} \sum_{t=0}^{J-1} \mathrm{PC}_{\mathrm{tot}}(t) = \mathrm{PC}_{\mathrm{tot}}^{\mathrm{av}} \tag{46}$$

$$\lim_{J\to\infty} \frac{1}{J} \sum_{t=0}^{J-1} R_{\mathrm{tot}}(t) = R_{\mathrm{tot}}^{\mathrm{av}} \tag{47}$$

$$\lim_{t\to\infty} \eta_{\mathrm{EE}}(t) = \frac{R_{\mathrm{tot}}^{\mathrm{av}}}{\mathrm{PC}_{\mathrm{tot}}^{\mathrm{av}}} = \eta_{\mathrm{EE}}^{\mathrm{av}}. \tag{48}$$

Under the assumptions (21)–(23) and (46)–(48), the equivalence of the following equations:

$$\begin{aligned} & \overline{\mathrm{PC}}_{\mathrm{tot}} = \mathrm{PC}_{\mathrm{tot}}^{\mathrm{av}},\ \overline{R}_{\mathrm{tot}} = R_{\mathrm{tot}}^{\mathrm{av}} \\ & \lim_{t\to\infty} \mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\right\} = \eta_{\mathrm{EE}}^{\mathrm{av}} = \eta_{\mathrm{EE}} \end{aligned} \tag{49}$$

is guaranteed by the Lebesgue dominated convergence theorem [34], [38], [41].

From (49), we can obtain

$$\lim_{J\to\infty} \mathbb{E}\left\{\frac{1}{J}\sum_{t=0}^{J-1}\left[\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t)\right]\right\} = \eta_{\mathrm{EE}}^{\mathrm{av}}\mathrm{PC}_{\mathrm{tot}}^{\mathrm{av}} = R_{\mathrm{tot}}^{\mathrm{av}} = \overline{R}_{\mathrm{tot}} \tag{50}$$

$$\lim_{J\to\infty} \frac{1}{J} \sum_{t=0}^{J-1} \mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\right\} = \eta_{\mathrm{EE}}^{\mathrm{av}}. \tag{51}$$

In addition, we denote the capacity region of the network by Λ, which is defined as the set of all the traffic arrival rates that can be stably supported by the network [42]. In other words,

there exists at least a resource allocation policy to stabilize the network under this arrival rate.

B. Analytical Results

With the help of Lemma 1 and the above preliminary, we can quantify the performance of the eTrans as follows, which is the question left in Section III-C.

Theorem 5: Suppose that problem (15) is feasible and $\mathbb{E}\{L(\boldsymbol{Q}(0))\} < \infty$. If $\boldsymbol{\lambda}$ is strictly interior to the network capacity region Λ, then the eTrans with any $V > 0$ has the following properties.

a) The performance bound of EE satisfies

$$\eta_{\mathrm{EE}} \geq \eta_{\mathrm{EE}}^{\mathrm{opt}} - \frac{\frac{B}{PC_{\min}}}{V}. \tag{52}$$

b) The average queue length $\overline{Q}$ has the following performance bound:

$$\begin{aligned}\overline{Q} &= \lim_{J\to\infty} \frac{1}{J} \sum_{t=0}^{J-1} \sum_{k=1}^{K} \mathbb{E}\left\{Q_k(t)\right\} \\ &\leq \frac{B + V\left[R_{\max} + \eta_{\mathrm{EE}}^{\mathrm{opt}}(\mathrm{PC}_{\max} - \mathrm{PC}_{\min})\right]}{\varepsilon}.\end{aligned} \tag{53}$$

Proof: Define an i.i.d. algorithm as the one that chooses a resource allocation strategy $\boldsymbol{\rho}(t)$, $\boldsymbol{P}(t)$, and $\boldsymbol{X}(t)$ independently and probabilistically in a predefined policy space according to a certain distribution in all slots t. For ease of understanding, we first review a basic result in the following lemma in the Lyapunov optimization technique. Its proof uses a standard result in the stochastic optimization theory [34].

Lemma 3: Suppose that $\boldsymbol{\lambda}$ is strictly interior to the capacity region Λ, and that $\boldsymbol{\lambda} + \varepsilon$ is also in Λ for a positive ε. Furthermore, (15) is feasible and the boundedness assumptions (21)–(23) hold. Then, for any $\delta > 0$, there exists an i.i.d. algorithm $\boldsymbol{\rho}(t)$, $\boldsymbol{P}(t)$, and $\boldsymbol{X}(t)$ that satisfies

$$\mathbb{E}\left\{R_{\mathrm{tot}}^*(t)\right\} \geq \mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}^*(t)\right\}\left(\eta_{\mathrm{EE}}^{\mathrm{opt}} - \delta\right) \tag{54}$$

$$\mathbb{E}\left\{R_k^*(t)|\boldsymbol{Q}(t)\right\} = \mathbb{E}\left\{R_k^*(t)\right\} \geq \lambda_k + \varepsilon \tag{55}$$

where $\mathrm{PC}_{\mathrm{tot}}^*(t)$ and $R_{\mathrm{tot}}^*(t)$ are the resulting values under i.i.d. $\boldsymbol{\rho}(t)$, $\boldsymbol{P}(t)$, and $\boldsymbol{X}(t)$.

Since the eTrans minimizes the right-hand side of (27), we have

$$\begin{aligned}&\Delta\left(\boldsymbol{Q}(t)\right) + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)|\boldsymbol{Q}(t)\right\} \\ &\leq B + \sum_{k=1}^{K} Q_k(t)\mathbb{E}\left\{A_k(t) - {R_k}^*(t)|\boldsymbol{Q}(t)\right\} \\ &\quad + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}^*(t) - R_{\mathrm{tot}}^*(t)|\boldsymbol{Q}(t)\right\}\end{aligned} \tag{56}$$

where $R_{\mathrm{tot}}^*(t)$ and $\mathrm{PC}_{\mathrm{tot}}^*(t)$ are the resulting values under any alternative (possibly i.i.d.) resource allocation policy $\boldsymbol{\rho}^*(t)$, $\boldsymbol{P}^*(t)$, and $\boldsymbol{X}^*(t)$.

Plugging (54) and (55) into (56) and taking a limit as $\delta \to 0$ yield

$$\begin{aligned}&\Delta\left(\boldsymbol{Q}(t)\right) + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)|\boldsymbol{Q}(t)\right\} \\ &\leq B - \varepsilon\sum_{k=1}^{K} Q_k(t) + V\eta_{\mathrm{EE}}(t)\mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}^*(t)\right\} - V\eta_{\mathrm{EE}}^{\mathrm{opt}}\mathbb{E}\{\mathrm{PC}_{\mathrm{tot}}^*(t)\}.\end{aligned} \tag{57}$$

a) Taking iterated expectation at both sides of (57) yields

$$\begin{aligned}&\mathbb{E}\left\{L\left(\boldsymbol{Q}(t+1)\right)\right\} - \mathbb{E}\left\{L\left(\boldsymbol{Q}(t)\right)\right\} \\ &\quad + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t) - R_{\mathrm{tot}}(t)\right\} \\ &\leq B - \varepsilon\sum_{k=1}^{K}\mathbb{E}\left\{Q_k(t)\right\} + V\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\right\}\mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}^*(t)\right\} \\ &\quad - V\eta_{\mathrm{EE}}^{\mathrm{opt}}\mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}^*(t)\right\}.\end{aligned} \tag{58}$$

Using telescoping sums over $t \in \{0, 1, \ldots, J-1\}$ and exploiting the fact that $Q_k(t) \geq 0$, we get

$$\begin{aligned}&\mathbb{E}\left\{L\left(\boldsymbol{Q}(J)\right)\right\} - \mathbb{E}\left\{L\left(\boldsymbol{Q}(0)\right)\right\} \\ &\quad + V\left[\sum_{t=0}^{J-1}\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\mathrm{PC}_{\mathrm{tot}}(t)\right\} - \sum_{t=0}^{J-1}\mathbb{E}\left\{R_{\mathrm{tot}}(t)\right\}\right] \\ &\leq J\left[B - V\eta_{\mathrm{EE}}^{\mathrm{opt}}\mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}^*(t)\right\}\right] \\ &\quad + V\mathbb{E}\left\{\mathrm{PC}_{\mathrm{tot}}^*(t)\right\}\sum_{t=0}^{J-1}\mathbb{E}\left\{\eta_{\mathrm{EE}}(t)\right\}.\end{aligned} \tag{59}$$

The remaining proof for a) and b) is similar to that in [33, Th. 2 (b) and (c), p. 6]; we thus omit it for brevity. ■

Remark 6: Equations (52) and (53) together show a tradeoff of $[O(1/V), O(V)]$ between EE and queue length (i.e., delay). That is, we can tune the EE–delay performance via V, which will be further verified in the following by simulations.

VI. Simulation Results and Analysis

In this section, we take completely random and dynamic HWNs as an example to evaluate the performance of the eTrans.

A. Parameters Setting

During the process of communication and mobility of users, we assume that the following cases will occur (see Fig. 1): 1) no Wi-Fi AP; 2) under the coverage of one Wi-Fi AP; 3) under the overlap range of two Wi-Fi APs; and 4) under the overlap range of three Wi-Fi APs. Each of the four cases has a probability of 0.4, 0.3, 0.2, and 0.1, respectively. It is worthwhile to note that the CN always exists in all cases.

For simplicity of the simulations, we normalize the subcarrier bandwidth to 1, i.e., $W = 1$, and assume that the channel gain $\boldsymbol{G}^{\mathrm{BS}}(t) = (g_{k,n}(t))$ is i.i.d. over slots with $g_{k,n}(t)$ generated randomly in 20 equal probability states {0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}. We determine the transmit rate from APs to MTs by applying the rate adaption scheme based on the SNR threshold, as shown in Table I [32]. Each state in Table I from 0 to 54 Mb/s has a probability of 0.08, 0.08, 0.15, 0.15, 0.15, 0.15, 0.08, 0.08, and 0.08. We take the drain efficiency of 35% for the power amplifier in the BS, i.e.,

TABLE I
SNR VERSUS RATE

SNR range (dB)	Rate (Mbps)
> 24.56	54
$[24.05, 24.56)$	48
$[18.8, 24.05)$	36
$[17.04, 18.8)$	24
$[10.79, 17.04)$	18
$[9.03, 10.79)$	12
$[7.78, 9.03)$	9
$[6.02, 7.78)$	6
< 6.02	0

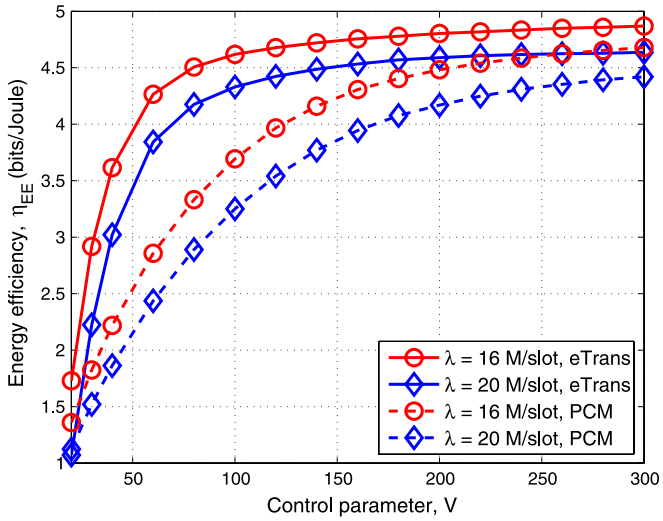


Fig. 2. EE versus control parameter V.

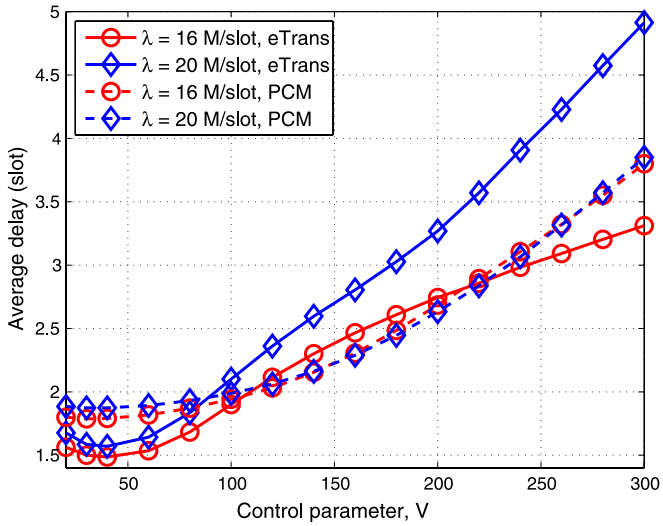


Fig. 3. Average delay versus control parameter V.

$1/\xi = 0.35$ [27], and set $K = 20$, $N = 256$, $P_{\text{tx}}^{\text{AP}} = 10.1$ W, and $P_{\text{idle}}^{\text{AP}} = 9.2$ W [23].

We simulate the eTrans for different control parameters V. Each point of the following curves is run for 40 000 slots and averaged over these values.

B. Quantitative EE–Delay Performance Control

In Fig. 2, the resulting EE is plotted against V. It is shown that η_{EE} converges as V increases. Hence, it inevitably converges to $\eta_{\text{EE}}^{\text{opt}}$ and can arbitrarily approach $\eta_{\text{EE}}^{\text{opt}}$ from (52).

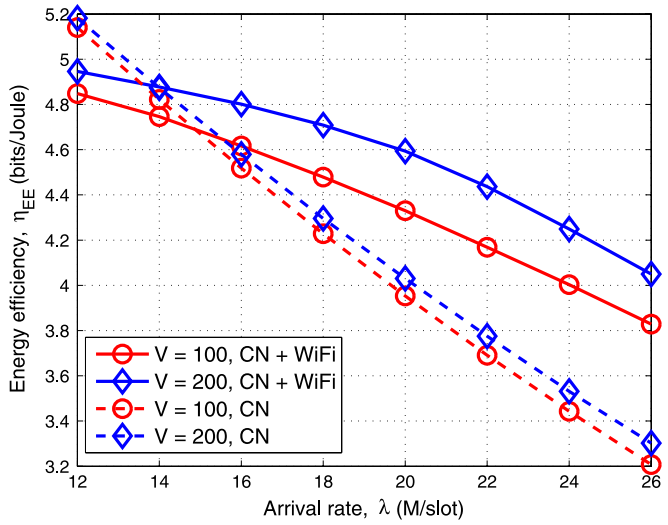


Fig. 4. EE versus traffic arrival rate $\boldsymbol{\lambda}$ under different control parameter V and scenarios.

Further, it is clear that EE increases to the optimal value at the speed of $O(1/V)$ as V increases for any given traffic arrival rate $\boldsymbol{\lambda} = (\lambda_k)$. Meanwhile, the average delay (i.e., queue length) grows linearly in $O(V)$, as shown in Fig. 3. Figs. 2 and 3 together show that there is a tradeoff between EE and delay, and it can be given by $[O(1/V), O(V)]$, which verifies the theoretical results (52) and (53) in Theorem 5. Hence, the eTrans provides an important method for the system to flexibly balance EE and average delay by simply adjusting control parameter V. Specifically, if the system prefers a better EE, a larger V is required (see Fig. 2). Otherwise, a smaller V is desired for a smaller delay (see Fig. 3).

In comparison, we also plot the curves of EE and delay produced by solving the PC minimization (PCM) problem [13], [14] but subject to the same constrains as in (15) (i.e., C1–C6). Note that [13], [14] assumed that subcarrier assignment, power allocation, and time fraction determination are all known parameters. As shown, these curves obtained from the eTrans and the PCM are distinct with each other. In other words, results from the power–delay tradeoff can hardly provide quantitative insights into EE–delay tradeoff issues. Hence, our proposed formulation and technique provide a significant means to investigate the EE–delay tradeoff.

C. Impacts of Parameters and Scenarios on System Performance

Figs. 4 and 5 display how the traffic arrival rate $\boldsymbol{\lambda}$, control parameter V, and scenarios affect the system EE and the average delay. Note that, in the figures, "CN + Wi-Fi" denotes scenarios that HWNs are composed of the CN and Wi-Fi, as shown in Fig. 1, whereas "CN" represents scenarios only consisting of the CN [33],[6] i.e., there is no Wi-Fi.

First, as can be seen, for any given V and scenario, both EE and the average delay deteriorate, i.e., EE decreases and the

[6]Note that, although with the same objective, the formulation to produce the curves for "CN" in Figs. 4 and 5 [cf. (15)] can be regarded as an extension of that in [33] because it jointly optimizes subcarrier assignment and power allocation instead of only power allocation in [33].

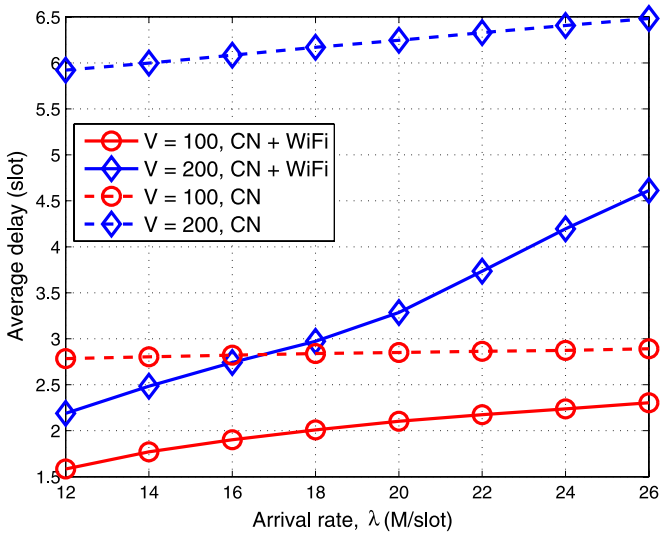


Fig. 5. Average delay versus traffic arrival rate $\boldsymbol{\lambda}$ under different control parameter V and scenarios.

average delay increases, with $\boldsymbol{\lambda}$. This is because it is required to increase the transmit rate to keep a finite queue length, i.e., to ensure the network stability [see C1 in (15)], which, naturally, is accompanied by the increased PC. However, an increase in the transmit power does not result in a proportional increase in the transmit rate due to the diminishing slope of the logarithmic rate-power function [see (1)]. As a result, EE decreases from (12).

Second, for a given traffic arrival rate $\boldsymbol{\lambda}$ and a specific scenario (e.g., CN + Wi-Fi), EE increases but the average delay degrades as V increases from 100 to 200. This follows from the fact that a larger V indicates the system emphasizes more on EE and less on the average queue length (i.e., the average delay) [see (29) and Remark 2]. This has also been theoretically proved in Theorem 5 [see (52) and (53)] and explicitly shown in Figs. 2 and 3.

Third, for a given traffic arrival rate $\boldsymbol{\lambda}$ and V, it is observed that different scenarios significantly affect the system performance. Specifically, "CN + Wi-Fi" outperforms "CN" in terms of both EE and the average delay. This is because the coexistence of the CN and Wi-Fi offers network diversity to MTs with multihoming capability. That is, MTs equipped with multiple interfaces are able to access multiple RANs simultaneously. Thus, it is more flexible for the system to allocate limited wireless resource to MTs as multihoming capability allows MTs to achieve their required QoS from all available RANs. In addition, Figs. 4 and 5 together indicate that HWNs improve system performance by jointly allocating resource across RANs.

VII. Conclusions

In this paper, we have formulated a stochastic optimization problem to investigate the delay-aware energy-efficient transmission problem in HWNs by jointly optimizing subcarrier assignment, power allocation, and time fraction determination. Resorting to the fractional programming theory and the Lyapunov optimization technique, we have devised the eTrans to solve the problem. Most importantly, we have developed the extremely simple and optimal algorithms for subcarrier assignment, power allocation, and time fraction determination with low complexity, where all of them have the closed-form solutions without requiring any iteration. The theoretical analysis and simulation results have verified the capability of the eTrans to flexibly control EE and average delay.

Appendix
Proof of Theorem 3

Assuming that $f(x)$ is concave, then its perspective function $tf(x/t)$ is still concave in (x,t) [39]. From this, $r_{k,n}(t) = \rho_{k,n}(t)W\log_2(1+(p_{k,n}(t)g_{k,n}(t)/\rho_{k,n}(t)))$ is jointly concave in $\boldsymbol{P}(t)$ and $\boldsymbol{\rho}(t)$ because it can be regarded as the perspective function of the concave function $\log_2(1+p_{k,n}g_{k,n})$. As a result, the objective in (38) is jointly convex in $\boldsymbol{P}(t)$ and $\boldsymbol{\rho}(t)$ as it is the sum of $K \times N$ convex functions.

In addition, C2, C4, and C6 in (38) are all linear constraints; thus, the sets produced by them for $\boldsymbol{P}(t)$ and $\boldsymbol{\rho}(t)$ are convex, respectively. Consequently, C2, C4, and C6 together construct a convex set as well.

Therefore, (38) is a convex optimization problem, because it minimizes a convex function over a convex set.

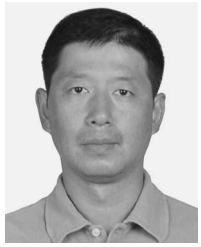
References

[1] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.
[2] J. Rao and A. Fapojuwo, "A survey of energy efficient resource management techniques for multicell cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 154–180, 1st Quart. 2014.
[3] C.-X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.
[4] L. Wang and G.-S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks–A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 271–292, Jan. 2013.
[5] A. Ahmed, L. Boulahia, and D. Gaïti, "Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 776–811, 2nd Quart. 2014.
[6] "The 1000× data challenge," Qualcomm, San Diego, CA, USA. [Online]. Available: http://www.qualcomm.com/solutions/wireless-networks/technologies/1000x-data
[7] M. Ismail and W. Zhuang, "A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 425–432, Feb. 2012.
[8] X. Pei, T. Jiang, D. Qu, G. Zhu, and J. Liu, "Radio-resource management and access-control mechanism based on a novel economic model in heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3047–3056, Jul. 2010.
[9] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 998–1006.
[10] Y. Choi, H. Kim, S. W. Han, and Y. Han, "Joint resource allocation for parallel multi-radio access in heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3324–3329, Nov. 2010.
[11] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim, "Radio resource management with proportional rate constraint in the heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1066–1075, Mar. 2012.
[12] M. Ismail, W. Zhuang, and S. Elhedhli, "Energy and content aware multihoming video transmission in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3600–3610, Jul. 2013.
[13] M.-R. Ra *et al.*, "Energy-delay tradeoffs in smartphone applications," in *Proc. MobiSys*, New York, NY, USA, Jun. 2010, pp. 255–270.
[14] P. Shu *et al.*, "eTime: Energy-efficient transmission between cloud and mobile devices," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 14–19.
[15] C. Li, J. Zhang, and K. Letaief, "Throughput and energy efficiency analysis of small cell networks with multi-antenna base stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2505–2517, May 2014.

[16] X. Ge *et al.*, "Spectrum and energy efficiency evaluation of two-tier femtocell networks with partially open channels," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1306–1319, Mar. 2014.
[17] J. Kwak, K. Son, Y. Yi, and S. Chong, "Greening effect of spatio-temporal power sharing policies in cellular networks with energy constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4405–4415, Dec. 2012.
[18] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 1–14, Jan. 2015.
[19] G. Lim, C. Xiong, L. Cimini, and G. Li, "Energy-efficient resource allocation for OFDMA-based multi-RAT networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2696–2705, May 2014.
[20] S. Kim, B. Lee, and D. Park, "Energy-per-bit minimized radio resource allocation in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1–12, Feb. 2014.
[21] X. Ma, M. Sheng, and Y. Zhang, "Green communications with network cooperation: A concurrent transmission approach," *IEEE Commun. Lett.*, vol. 16, no. 12, pp. 1952–1955, Dec. 2012.
[22] M. Ismail and W. Zhuang, "Green radio communications in a heterogeneous wireless medium," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 128–135, Jun. 2014.
[23] S. Kim, S. Choi, and B. G. Lee, "A joint algorithm for base station operation and user association in heterogeneous networks," *IEEE Commun. Lett.*, vol. 17, no. 8, pp. 1552–1555, Aug. 2013.
[24] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.
[25] Z. Shen, J. Andrews, and B. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.
[26] D. Ng, E. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.
[27] C. Xiong, G. Li, S. Zhang, Y. Chen, and S. Xu, "Energy- and spectral-efficiency tradeoff in downlink OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3874–3886, Nov. 2011.
[28] Y. Li *et al.*, "Energy-efficient subcarrier assignment and power allocation in OFDMA systems with max-min fairness guarantees," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3183–3195, Sep. 2015.
[29] Y. Li, M. Sheng, X. Wang, Y. Zhang, and J. Wen, "Max-min energy-efficient power allocation in interference-limited wireless networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 9, pp. 4321–4326, Sep. 2015.
[30] Y. Li, M. Sheng, C. Yang, and X. Wang, "Energy efficiency and spectral efficiency tradeoff in interference-limited wireless networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1924–1927, Oct. 2013.
[31] J. Choi, J. Yoo, S. Choi, and C. Kim, "EBA: An enhancement of the IEEE 802.11 DCF via distributed reservation," *IEEE Trans. Mobile Comput.*, vol. 4, no. 4, pp. 378–390, Jul. 2005.
[32] J. Yee and H. Pezeshki-Esfahani, "Understanding wireless LAN performance trade-offs," *Commun. Syst. Design*, vol. 11, pp. 32–35, Nov. 2002.
[33] Y. Li, M. Sheng, Y. Shi, X. Ma, and W. Jiao, "Energy efficiency and delay tradeoff for time-varying and interference-free wireless networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 5921–5931, Nov. 2014.
[34] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Vermont, Vic., Australia: Morgan & Claypool, 2010.
[35] Y. Li *et al.*, "Throughput-delay tradeoff in interference-free wireless networks with guaranteed energy efficiency," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1608–1621, Mar. 2015.
[36] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1987.
[37] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.
[38] M. J. Neely, "Dynamic optimization and learning for renewal systems," *IEEE Trans. Autom. Control*, vol. 58, no. 1, pp. 32–46, Jan. 2013.
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[40] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab Software for Disciplined Convex Programming," v. 2.0 beta, Sep. 2012. [Online]. Available: http://cvxr.com/cvx
[41] D. Williams, *Probability With Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[42] M. J. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2915–2934, Jul. 2006.

Yuzhou Li received the B.Eng. degree in electronic and information engineering from Jilin University, Changchun, China, in 2009. He is currently working toward the Ph.D. degree with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an, China.

His research interests include green communications, wireless resource allocation, and convex optimization and stochastic network optimization and their applications in wireless communications.



Yan Shi (M'10) received the Ph.D. degree from Xidian University, Xi'an, China, in 2005.

He is currently an Associate Professor with the State Key Laboratory of Integrated Service Networks, Xidian University. His current research interests include cognitive networks, modern switching technologies, and distributed wireless networking.

Min Sheng (M'03) received the M.Eng and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1997 and 2000, respectively.

She is currently a Full Professor with the Broadband Wireless Communications Laboratory and the State Key Laboratory of Integrated Service Networks, School of Telecommunication Engineering, Xidian University. She is the author of two books and over 50 papers in refereed journals and conference proceedings. Her main research interests include mobile ad hoc networks, wireless sensor networks, wireless mesh networks, third-generation/fourth-generation mobile communication systems, dynamic radio resource management for integrated services, cross-layer algorithm design and performance evaluation, cognitive radio and networks, cooperative communications, and medium access control protocols.

Dr. Sheng received the New Century Excellent Talents in University from the Ministry of Education of China and the Young Teachers Award from the Fok Ying-Tong Education Foundation, China, in 2008.

Cheng-Xiang Wang (S'01–M'05–SM'08) received the B.Sc. and M.Eng. degrees in communication and information systems from Shandong University, Jinan, China, in 1997 and 2000, respectively, and the Ph.D. degree in wireless communications from Aalborg University, Aalborg, Denmark, in 2004.

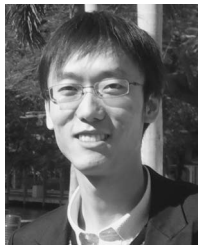
From 2000 to 2001, he was a Research Assistant with the Hamburg University of Technology, Hamburg, Germany. From 2001 to 2005, he was a Research Fellow with the University of Agder, Grimstad, Norway. In 2004, he was a Visiting Researcher with Siemens AG-Mobile Phones, Munich, Germany. Since 2005, he has been with Heriot-Watt University, Edinburgh, U.K., where he was promoted to Professor in 2011. He is also an Honorary Fellow of the University of Edinburgh and a Chair/Guest Professor with both Shandong University and Southeast University, Nanjing, China. He is the author of one book chapter and over 200 papers in refereed journals and conference proceedings, as well as the editor of one book. His research interests include wireless channel modeling and simulation, green communications, cognitive radio networks, vehicular communication networks, massive multiple-input–multiple-output systems, and fifth-generation wireless communications.

Dr. Wang has served as a Technical Program Committee (TPC) Member, a TPC Chair, and a General Chair for more than 70 international conferences. He was the lead Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS special issue on vehicular communications and networks. He has served as an Editor for eight international journals, including the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (since 2011) and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2007–2009). He received Best Paper Awards from the 2010 IEEE Global Communications Conference, the 2011 IEEE International Conference on Communication Technology, the 2012 IEEE International Conference on ITS Telecommunications, and the 2013 Spring IEEE Vehicular Technology Conference. He is a Fellow of the Institution of Engineering and Technology and the Higher Education Academy and a member of the Engineering and Physical Research Council Peer Review College.

Jiandong Li (SM'05) received the B.E., M.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 1982, 1985, and 1991 respectively, all in communications engineering.

Since 1985, he has been a faculty member with the School of Telecommunications Engineering, Xidian University, where he is currently a Professor and the Vice Director of the academic committee of the State Key Laboratory of Integrated Service Networks. From 2002 to 2003, he was a Visiting Professor with the Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. His main research interests include communications and information systems, cognitive radio, and signal processing.

Dr. Li served as the General Vice Chair for the 2009 International Conference on Communications and Networking and as the Technical Program Committee Chair for the 2013 IEEE/CIC International Conference on Communications in China. He received the Distinguished Young Researcher Award from the National Natural Science Foundation of China and the Changjiang Scholar Award from the Ministry of Education, China.



Xijun Wang (M'12) received the B.S. degree (with distinction) in telecommunications engineering from Xidian University, Xi'an, China, in 2005 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in January 2012.

Since 2012, he has been with School of Telecommunications Engineering, Xidian University, where he is currently an Assistant Professor. His research interests include wireless communications, cognitive radios, and interference management.

Dr. Wang served as a Publicity Chair for the 2013 IEEE/CIC International Conference on Communications in China. He received the "Outstanding Graduate of Shaanxi Province" Award in 2005, the Excellent Paper Award at the Sixth International Student Conference on Advanced Science and Technology in 2011, and the Best Paper Award at IEEE/CIC ICCC in 2013.

Yan Zhang (M'12) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2005 and 2010, respectively.

He is currently an Associate Professor with Xidian University. His research interests include cooperative cognitive networks, self-organizing networks, media access protocol design, energy-efficient transmission, and dynamic radio resource management in heterogeneous networks.