IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 3, 2018, accepted March 4, 2018, date of publication April 19, 2018, date of current version May 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2823004

QoS-Aware Utility-Based Resource Allocation in Mixed-Traffic Multi-User OFDM Systems

SUMAN KHAKUREL[1], LEILA MUSAVIAN[2], (Member, IEEE), HUNG V. VU[3], (Student Member, IEEE), AND THO LE-NGOC[3], (Fellow, IEEE)
[1]Index Exchange, North York, ON M6B 1P5, Canada
[2]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K.
[3]Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada

Corresponding author: Hung V. Vu (hung.vu2@mail.mcgill.ca)

This work was supported in part by the U.K. EPSRC under Grant EP/N032268/1 and in part by the Natural Sciences and Engineering Research Council of Canada.

ABSTRACT This paper deals with the joint subcarrier and power allocation problem in a downlink multi-user orthogonal frequency division multiplexing system subject to user delay and minimum rate quality-of-service (QoS) requirements over a frequency-selective multi-carrier fading channel. We aim to maximize the utility-pricing function, formulated as the difference between the achieved spectral efficiency and the associated linear cost function of transmit power scaled by a system-dependent parameter. For a homogeneous system, we show that the joint resource allocation can be broken down into sequential problems while retaining the optimality. Specifically, the optimal solution is obtained by first assigning each subcarrier to the user with the best channel gain. Subsequently, the transmit power for each subcarrier is adapted according to water-filling policy if the global optimum is feasible, else it is given by a non-water-filling power adaptation. For a heterogeneous system, an optimal solution needs exhaustive search and hence, we resort to two reduced-complexity sub-optimal algorithms. Algorithm-I is a simple extension of the aforementioned optimal algorithm developed for a homogeneous system, while Algorithm-II further takes into consideration the heterogeneity in user QoS requirements for performance enhancement. Simulation results reveal the impacts of user QoS requirements, number of subcarriers and number of users on the system transmit power.

INDEX TERMS Algorithms, effective capacity, OFDM, QoS, resource allocation, utility-pricing function.

I. INTRODUCTION

Next-generation wireless communication systems are expected to support traffic demanding diverse levels of quality-of-service (QoS) requirements [1]. These communication systems need to employ efficient and feasible mechanisms for guaranteeing QoS such as, resource reservation and admission control in order to meet this expectation. To design effective mechanisms, simple and precise modeling of wireless channels in terms of QoS metrics (e.g., delay, delay-outage probability, and data rate) are necessary [1], [2]. This has laid the path for a cross-layer analysis of the wireless fading channel through a link-layer channel model called effective capacity (EC) [2]. EC is a QoS-aware metric that determines the maximum constant arrival rate which a system can support while satisfying a target delay requirement specified by a delay-QoS exponent. By using the EC model, a rate-efficient power allocation technique under delay-outage probability constraint for a single-channel flat-fading system is proposed in [3]. This is further extended to frequency-selective multi-carrier system in [4], where it is shown that the EC decreases from ergodic capacity to zero-outage-capacity as the QoS exponent increases from zero to infinity. An optimal strategy, with respect to joint subcarrier and power allocation, for maximizing the EC of a multi-user orthogonal frequency division multiple access (OFDMA) relay-based networks under average power constraint is developed in [5].

With the increasing demand for higher capacity links, the device power consumption is also increasing [6]. Thus, increasing the energy efficiency (EE) in cellular networks has become an urgent need. Maximizing the EE of a system without delay constraint has been studied for various transmission scenarios, e.g., point-to-point communication link [7], single-user orthogonal frequency division multiplexing (OFDM) systems [8], [9] and multi-user OFDM systems [10]–[12], wherein, EE, in b/J/Hz, is defined as the ratio of Shannon capacity to the total dissipated power.

Since EC and EE are two important performance metrics, joint-study of EC and EE is necessary for designing future communication systems. For instance, improving EE to reduce the power consumption is the key factor to maintain sustainable green networks [13], [14]. By employing the EC formulation, instead of Shannon capacity, investigation of EE under delay-QoS constraint for flat-fading channels has been studied in [15]. In particular, an optimal power allocation technique for maximizing the EE of a delay limited system over flat-fading channels is proposed. Instead of optimizing EE, Gurosy *et al.* [16] determine the minimum bit energy required to satisfy a specified delay constraint in the low-power and wideband regimes. They also analyze the rate-energy trade-off of flat-fading channels under delay-outage probability constraint. The problem of transmit power minimization for a multi-user multiple-input multiple-output (MIMO) time division multiple access (TDMA) system is considered in [17] subject to constraints on the EC of individual user. A closed-form expression for the EC of a single-user MIMO system is derived, which is used as a basis for dynamic time sharing and power allocation in multi-user MIMO systems. Wu *et al.* [18] also investigate the joint transmitter and receiver optimization for the EE in OFDMA systems but does not consider the delay constraint, which is an important factor in communication systems. Phan *et al.* [19] develop an online scheduling algorithm for meeting the minimum rates of users with statistical delay guarantees in a multi-user OFDM system while minimizing the total power of the system. A low-complexity heuristic algorithm is proposed for utility-based resource allocation in a wireless network with mixed traffic in [20]. Utility function for best-effort (BE) traffic is modeled as a convex function while that for real-time (RT) traffic is modeled as a sigmoid function with respect to the allocated resource. The unified utility function is then defined as an aggregate utility of all the users in the network. The proposed algorithm guarantees the QoS requirement for the RT traffic preferentially and makes a tradeoff between throughput and fairness for users with BE traffic. In [21], employing utility-based framework, a polynomial-time heuristic algorithm is proposed to balance spectral efficiency (SE) and fairness of multi-user OFDM system with mixed RT and non-real-time (NRT) traffic patterns, while satisfying QoS requirements of RT users. Utility function of each user is modeled as an increasing function of wireless link quality and data rate for NRT user while wireless link quality, data rate and average experienced delay for RT user. The total network utility, represented as the summation of the utilities of all served users at each time, is then maximized with the overall computational complexity order of $O(U^2N^2)$ where U and N represent the total number of users and subcarriers in the system, respectively. We observe that limited literature is available on the performance study of a mixed-traffic multi-user energy-efficient OFDM system, where each user carries information corresponding to both BE and QoS-aware traffic. This study is important because energy is a valuable commodity and mixed-traffic scenario is a very realistic traffic model for future Long Term Evolution (LTE) networks which is likely to carry both guaranteed bit rate (GBR) and default bearer/non-GBR traffic. A QoS-aware or GBR traffic, like voice over Internet Protocol (VoIP), requires a fixed data-rate and/or delay requirement while BE or default bearer traffic, like data traffic, consumes the leftover capacity not used by the GBR sources [22].

In this paper, we consider the joint subcarrier and power allocation problem in a downlink (DL) multi-user OFDM system subject to constraints on the EC of QoS-aware applications, co-exisiting with BE applications, for each user.[1] The adaptive resource allocation aims to maximize the utility-pricing function, in b/s/Hz, formulated as the difference between the SE and the transmit power scaled by a parameter β, representing the penalty on transmit power. A higher value of β corresponds to higher penalty in transmit power. This function can be interpreted as a measure of energy-awareness in the system. We consider cases with homogeneity and heterogeneity in the system. A system is homogeneous if all the users in the system have same requirement on minimum EC and delay-QoS exponent corresponding to QoS-aware traffic, else it is heterogeneous. We show that, in a homogeneous system, optimal solution can be obtained by first allocating each subcarrier to the user with the best channel gain over that subcarrier and then adapting the transmit power for each allocated user-subcarrier pair. The transmit power is adapted according to water-filling policy if the global optimum is feasible, else it is given by a non-water-filling power adaptation. This algorithm is no longer optimal for heterogeneous systems due to heterogeneity in QoS requirements of users and we resort to two reduced-complexity sub-optimal algorithms: Algorithm-I is a simple extension of the aforementioned optimal algorithm developed for a homogeneous system, while Algorithm-II further takes into consideration the heterogeneity in user QoS requirements for performance enhancement. The complexity of these algorithms is of the order $O(UN)$. This is much less compared to the complexity of the optimal exhaustive-search algorithm which is of the order $O(NU^N)$.

We perform simulations to observe the impact of delay-QoS exponent (θ), minimum EC (μ^{min}), parameter (β), number of subcarriers (N) and number of users (U) on the transmit power. Simulation results indicate that, for the lower values, transmit power is indifferent to the variation in μ^{min}; however, it increases exponentially with further increment in μ^{min}. Moreover, transmit power exponentially decays as a function of β. For higher β, the maximization of utility-pricing function is transformed into power minimization problem and the transmit power is then invariant to β. Further investigation shows that, for higher β and μ^{min}, BE traffic cannot be

[1] A single-user system is studied in the conference version of this paper in [23].

supported by the system. For systems with higher values of β, total power is the minimum power required to meet the QoS requirements of users and it exponentially decreases with N due to increment in system bandwidth and frequency diversity. However, for the systems with smaller β, it is an increasing function of N. Next, performance comparison of the proposed sub-optimal algorithms with the optimal algorithm, for a heterogeneous system, shows that they provide near-optimal solutions. Moreover, Algorithm-II outperforms Algorithm-I whenever the globally optimal solution is infeasible, else they behave in a similar way.

The rest of the paper is organized as follows. Section II presents the system modeling and the formulation of optimization problem. The optimal and sub-optimal algorithms for subcarrier allocation and power adaptation for homogeneous and heterogeneous systems are provided in III. Section IV includes the numerical results. Finally, we draw conclusions in Section V.

II. PROBLEM FORMULATION

A. SYSTEM MODEL

As demonstrated in Fig. 1, we consider the DL transmission scenario of a single-cell multi-user OFDM system with one base station (BS) at the center of the cell and U users uniformly distributed across the cell. We consider N subcarriers, each with bandwidth b, in the cell and thus, the total bandwidth of the system, B, is Nb. These subcarriers are to be scheduled among U users in the beginning of each scheduling interval. Moreover, the channel state information (CSI) corresponding to each subcarrier is perfectly estimated at the receiver and reliably fed back to the transmitter, e.g. by using a pilot-based technique. When the CSI is available, the transmitter can adapt its transmit power according to the channel variation. Further, we assume a block-fading channel model [24], wherein, the channel condition remains constant over the duration of a fading block and independently changes from one block to another.

Let $\mathcal{N} = \{1, 2, \ldots, N\}$ be the set of subcarriers, $\mathcal{U} = \{1, 2, \ldots, U\}$ be the set of users and $h^t_{u,n}$ be the channel amplitude of user $u \in \mathcal{U}$ on subcarrier $n \in \mathcal{N}$ in slot $t = 1, 2, \ldots$. We assume that the channel gains, $h^t_{u,n}, \forall u, n, t$, are independent and identically distributed (i.i.d.) with some probability density function (PDF) f. If $p^t_{u,n}$ is the transmission power of user u over subcarrier n in slot t then the maximum SE of the user over that subcarrier in that slot is given by:

$$r^t_{u,n} = \log_2(1 + \gamma^t_{u,n} p^t_{u,n}), \tag{1}$$

where $\gamma^t_{u,n} = |h^t_{u,n}|^2 / Pl_u N_0 b$ is the CNR of user u on subcarrier n in slot t. Here, Pl_u is the distance-based path loss between user u and the BS and N_0 is the noise spectral density.

For QoS-aware applications, user traffic must satisfy a statistical delay constraint. In particular, assuming that the steady-state queue length exists, it is required that the probability for the queue length of the user u at equilibrium, $Q_u(\infty)$, exceeding a certain threshold for that user, x_u, exponentially decays as a function of x_u. This exponential decay can be characterized by a delay exponent θ_u [2]:

$$\theta_u = - \lim_{x_u \to \infty} \frac{\ln(Pr\{Q_u(\infty) > x_u\})}{x_u}, \tag{2}$$

where $Pr\{y > z\}$ denotes the probability that the inequality $y > z$ holds. The large and small values of θ_u correspond to fast and slow decaying rates indicating stringent and loose delay requirements, respectively. The application corresponding to user u can tolerate arbitrarily long delay when $\theta_u \to 0$, while it cannot tolerate any delay when $\theta_u \to \infty$. Further, the probability that the delay exceeds the maximum delay bound, D^{max}_u, corresponding to user u, defined as delay-outage probability, can be approximated as [1]:

$$P^{out}_u = Pr\{\text{Delay} > D^{max}_u\} \simeq \phi_u e^{-\mu_u \theta_u D^{max}_u}. \tag{3}$$

For a given constant arrival rate μ_u, ϕ_u is the probability that the buffer of user u is non-empty and can be approximated as the ratio of the constant arrival rate to the average service rate [1], i.e., $\phi_u = \frac{\mu_u}{\mathbb{E}\{R_u(k)\}}$, where $R_u(k)$ denotes the instantaneous discrete-time stationary and ergodic service process of user u at time k and $\mathbb{E}\{.\}$ denotes the statistical expectation. Let us define $\tilde{S}_u(t) = \sum_{k=1}^{t} R_u(k)$ as the service provided by the channel. Given the assumptions for the Gartner-Ellis theorem [25] are satisfied where Gartner-Ellis limit of $\tilde{S}_u(t)$, expressed as $\Lambda_u(\theta_u) = \lim_{t\to\infty} \frac{1}{t} \ln\left(\mathbb{E}\{e^{-\theta_u \tilde{S}_u(t)}\}\right)$, exists and is a convex function differentiable for all real $\theta_u \geq 0$ [2], the EC function of $\tilde{S}_u(t)$ can then be defined as [1]:

$$\mu_u = \frac{-\lim_{t\to\infty} \frac{1}{t} \ln\left(\mathbb{E}\{e^{-\theta_u \tilde{S}_u(t)}\}\right)}{\theta_u}, \quad \forall \theta_u \geq 0. \tag{4}$$

When the service process of user u $\{R_u(k), k = 1, 2, \ldots\}$ is uncorrelated, the EC expression in (4) can be reduced to:

$$\mu_u(\theta_u) = -\frac{1}{\theta_u} \ln\left(\mathbb{E}\{e^{-\theta_u R_u(k)}\}\right). \tag{5}$$

Normalizing (5) to the fading-block length T_f and the system bandwidth B, the EC of i.i.d block fading subcarriers corresponding to user u, in b/s/Hz, can be expressed as:

$$\mu_u(\theta_u) \triangleq -\frac{1}{\theta_u T_f B} \ln\left(\mathbb{E}\{e^{-\theta_u T_f \frac{B}{N} R_u}\}\right). \tag{6}$$

By the definition of EC, $\mu_u(\theta_u)$ is the maximum constant arrival rate, in b/s/Hz, corresponding to user u that the system can support in order to guarantee the statistical delay requirement defined by θ_u.

The issues on efficient resource allocation have been well studied in economics, where utility-pricing functions are used to quantify the benefit and cost of usage of certain resources. The basic idea of utility-pricing structures is to map the resource use (bandwidth, power, etc.) or performance criteria (data rate, delay, etc.) into the corresponding utility or price values and optimize the established utility-pricing system [26]. In this paper, we define utility-pricing function of the system (η) as the difference between the rate r and a function of power p representing the cost paid to achieve that rate. Mathematically, $\eta = r - c(p)$, where $c(p)$ is the cost function. As aforementioned, power is itself a valuable commodity and hence, we use cost function to reflect the expenses of power consumption. It is mentioned in [27] that there are at least two requirements for a cost function, viz., $c(0) = 0$ and $c(p)$ increases in power p. Alike [27], in this paper, we will use a linear cost function, i.e., $c(p) = \beta p$, where the parameter β is the pricing coefficient, in b/s/Hz/W, and is a constant independent of p. However, β is a system-dependent parameter. The system is power constrained if β is set to higher values, i.e., $\beta \gg 1$, and $\eta \approxeq -\beta p$ for those systems. When $\beta \to 0$, utility-pricing function is equivalent to the SE of the system, i.e., $\eta \approxeq r$. Further, the time-average utility-pricing function of the system, η^{av}, is given as[2]:

$$\eta^{av} \triangleq \lim_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T}\sum_{u=1}^{U}\sum_{n=1}^{N}\alpha_{u,n}^{t}(r_{u,n}^{t} - \beta p_{u,n}^{t})\right\}$$
$$= \mathbb{E}\left\{\sum_{u=1}^{U}\sum_{n=1}^{N}(\alpha_{u,n}(r_{u,n} - \beta p_{u,n}))\right\}, \quad (7a)$$

where $r_{u,n}^t$ and $p_{u,n}^t$ are the rate and power of user u over subcarrier n in time slot $t, \forall u \in \mathcal{U}, n \in \mathcal{N}$ and $\alpha_{u,n}^t \in \{0, 1\}$ is an indication variable representing the allocation of subcarrier n to user u in time slot t. $\alpha_{u,n}^t = 1$ if subcarrier n is assigned to user u in slot t and $\alpha_{u,n}^t = 0$, otherwise. This ensures OFDMA in the cell, meaning that a particular subcarrier in the cell can be used by at most one user in that cell at a given time. Also, $\mathbb{E}\{.\}$ denotes the statistical expectation with respect to the distribution f and (7a) holds true because the system is considered to be ergodic.

B. OPTIMIZATION PROBLEM

We aim to maximize the ergodic utility-pricing function of the system where each user has the data corresponding to QoS-aware and BE applications while maintaining the minimum EC for each user. Note that, meeting the minimum EC of user u is equivalent to meeting the delay requirement and minimum rate of the QoS-aware applications for that user specified by θ_u and μ_u, respectively. In each slot, we perform subcarrier scheduling among users and then determine the transmission power for each allocated subcarrier. Hence, the problem can be formulated as:

$$\text{maximize } \eta^{av}(\boldsymbol{\alpha}, P) = \mathbb{E}\left\{\sum_{u=1}^{U}\sum_{n=1}^{N}\alpha_{u,n}(r_{u,n} - \beta p_{u,n})\right\} \quad (8a)$$
$$\text{subject to } \mu_u(\theta_u) \geq \mu_u^{min}, \quad \forall u \in \mathcal{U}, \quad (8b)$$
$$\sum_{u=1}^{U}\alpha_{u,n} = 1, \quad \forall n \in \mathcal{N}, \quad (8c)$$
$$\alpha_{u,n} \in \{0, 1\}, \quad \forall u \in \mathcal{U},\ n \in \mathcal{N}, \quad (8d)$$
$$p_{u,n} \geq 0, \quad \forall u \in \mathcal{U},\ n \in \mathcal{N}. \quad (8e)$$

[2]Hereafter, for the ease of notation, we remove the time index t, whenever it is clear in the text.

Here, $\boldsymbol{\alpha}$ is a $U \times N$ matrix with each element, $\alpha_{u,n} \in \{0, 1\}$, indicating the allocation of subcarrier n to user u, P is a $U \times N$ matrix with each element, $p_{u,n}$, representing the allocated power to UE u on subcarrier n and μ_u^{min} is the minimum EC corresponding to user u. Constraints (8c) and (8d) enforce the OFDMA assumption in the cell.[3]

III. SUBCARRIER AND POWER ALLOCATION

Problem (8) is a combinatorial optimization problem with non-linear constraints and it may be difficult to obtain an optimal solution within any reasonable time frame [28]. Hence, we devise computationally-effective optimal or sub-optimal scheduling algorithms, whichever applicable, using the following propositions.

Proposition 1: The ergodic utility-pricing function of an OFDMA system with no constraints on the minimum EC of any user is maximum when each subcarrier is allocated to the user with the best channel gain on that subcarrier, i.e., $\alpha_{v_n,n} = 1$ and $\alpha_{u,n} = 0, \forall u \in \mathcal{U}\backslash\{v_n\}$, where $v_n = \text{argmax}_{u\in\mathcal{U}}\, \gamma_{u,n}, \forall n \in \mathcal{N}$.

Proof: Consider a single-user OFDM system with N subcarriers. Then, the maximization of unconstrained ergodic utility-pricing function of the system can be represented as:

$$\max_{p\in\mathbb{R}_+^N} \eta^{av}(p) = \mathbb{E}\left\{\sum_{n=1}^{N}\log_2(1+\gamma_n p_n) - \beta\sum_{n=1}^{N}p_n\right\}. \quad (9)$$

Problem (9) is concave in p. Hence, the Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient for optimality. At the optimal power allocation, p^*, we have

$$\left.\frac{\delta\eta^{av}}{\delta p_{v_n,n}}\right|_{p=p^*} = 0, \quad \forall n, \quad (10a)$$
$$\frac{\gamma_{v_n,n}}{(1+\gamma_{v_n,n}p_{v_n,n}^*)\ln 2} - \beta = 0, \quad \forall n, \quad (10b)$$
$$p_n^* = p_{v_n,n}^* = \left[\frac{1}{\beta\ln 2} - \frac{1}{\gamma_{v_n,n}}\right]^+, \quad \forall n. \quad (10c)$$

Here, $[x]^+ = \max\{0, x\}$. Let us assume that $p_n^* > 0, \forall n \in \mathcal{N}$. Substituting (10c) in (9), optimal utility-pricing function of the system can be computed as:

$$\eta^{av*} = \mathbb{E}\left\{\sum_{n=1}^{N}\log_2\left(\frac{\gamma_n}{\beta\ln 2}\right) - \frac{1}{\ln 2} + \frac{\beta}{\gamma_n}\right\}. \quad (11)$$

Thus, η^{av*}, is an increasing function of γ_n if $\frac{\gamma_n}{\beta\ln 2} > 1$, $\forall n \in \mathcal{N}$. If $\gamma_n \leq \beta\ln 2$ for any $n \in \mathcal{N}$ then $p_n^* = 0$. This implies that the unconstrained ergodic utility-pricing function (η^{av}) is an increasing function of channel amplitude

[3]The constraint (8e) only limits the value of the power to non-negative values and does not constraint the power on the upper side. We note that since the transmission power is used as a pricing factor in the objective function, the optimal value of the transmission power will not reach infinity. We assume that the transmit power can be big enough to achieve the optimal value of the optimization problem in (8). If the transmission power is limited to a maximum power, the total optimla power should be limited to the maximum power and the the solution of (8).

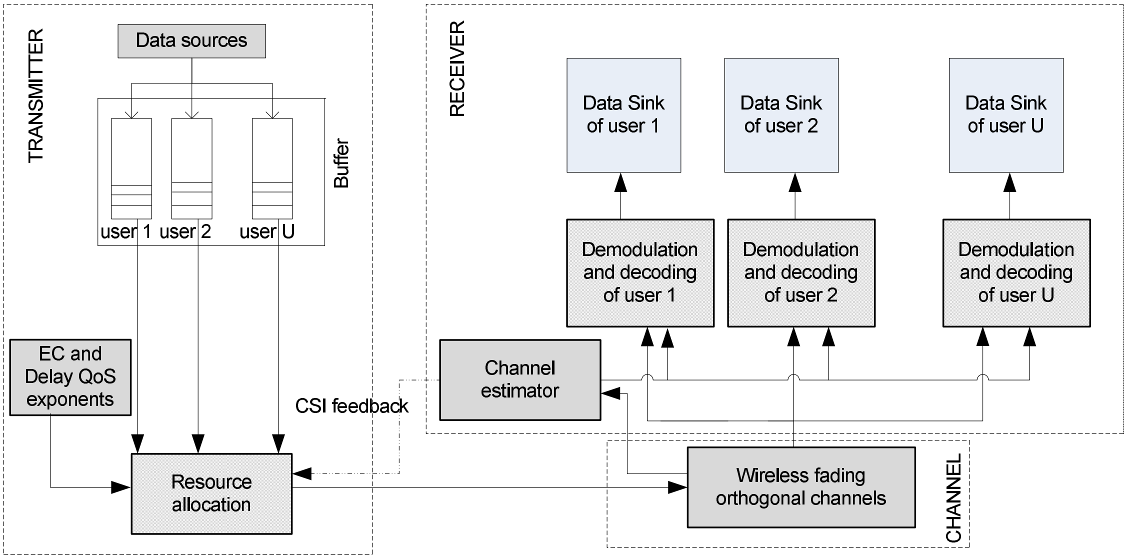

FIGURE 1. System model.

gains, γ_n, $\forall n \in \mathcal{N}$. Hence, η^{av*} in a multi-user OFDMA system with no constraints on the minimum EC of any user is obtained by allocating the subcarrier n to the user with the best channel gain. ■

Proposition 2: The optimal solution to (8) is either the global optimum or the boundary point.

Proof: Let $\boldsymbol{\alpha}_{arb}$ be any arbitrary subcarrier allocation matrix. After the subcarrier allocation, multi-user OFDMA system can be viewed as a single-user OFDM system. Hence, the problem of maximizing unconstrained ergodic utility-pricing function of the system is given by (9). At optimality,

$$p_n^* = \left[\frac{1}{\beta \ln 2} - \frac{1}{\gamma_n}\right]^+, \quad \forall n \in \mathcal{N} \text{ and} \tag{12}$$

$$\eta^{av*} = \mathbb{E}\left\{\sum_{n=1}^{N} \log_2\left(\frac{\gamma_n}{\beta \ln 2}\right) - \frac{1}{\ln 2} + \frac{\beta}{\gamma_n}\right\}. \tag{13}$$

Let $p' = p^* + \triangle p$, where p' and $p^* \in \mathbb{R}_+^N$ while $\triangle p \in \mathbb{R}^N$ represents a sufficiently small change in transmit power vector. Then, η^{av} at p' is given as $\mathbb{E}\left\{\sum_{n=1}^{N} \log_2\left(\frac{\gamma_n}{\beta \ln 2} + \gamma_n \triangle p_n\right) - \frac{1}{\ln 2} + \frac{\beta}{\gamma_n} - \beta \triangle p_n\right\}$. Differentiating η^{av} with respect to $\triangle p_n$, we get

$$\frac{\delta \eta^{av}}{\delta \triangle p_n} = \frac{1}{\ln 2} \times \frac{1}{1/(\beta \ln 2) + \triangle p_n} - \beta, \quad \forall n \in \mathcal{N}, \tag{14}$$

$$= \frac{\beta}{1 + \beta \triangle p_n \ln 2} - \beta, \quad \forall n \in \mathcal{N}. \tag{15}$$

Here, β is a non-negative parameter and $\ln 2 = 0.693 > 0$. If $\triangle p_n > 0$ then $\frac{\delta \eta^{av}}{\delta \triangle p_n}$ is negative and a decreasing function of $\triangle p_n, \forall n \in \mathcal{N}$. If $\triangle p_n < 0$ and is sufficiently small then $\frac{\delta \eta^{av}}{\delta \triangle p_n} > 0, \forall n \in \mathcal{N}$. Hence, $\eta^{av}(p)$, is an increasing function of the power vector p till global optimum. However, beyond optimality, ergodic utility-pricing function decreases with p. Indeed, taking the partial differentiation of (9) with respect to p_n yields the following inequality

$$\frac{\partial \eta^{av}(p)}{\partial p_n} = \frac{\gamma_n}{\ln 2(1 + \gamma_n p_n)} - \beta \leq 0, \quad \text{or}$$
$$p_n \geq \frac{1}{\beta \ln 2} - \frac{1}{\gamma_n}, \quad \forall n \in \mathcal{N}.$$

This implies that the optimal solution to (8) is either the global optimum, if feasible, or the boundary point of the function (8a) where it intersects with the constraint (8b). ■

We study solutions to (8) for both homogeneous and heterogeneous system. A system is considered to be homogeneous when all the users in the system have the same requirement on the minimum EC and the delay-QoS exponent corresponding to QoS-aware applications, *i.e.*, $\mu_u^{min} = \mu^{min}$ and $\theta_u = \theta, \forall u \in \mathcal{U}$. Whereas, in a heterogeneous system, users may have different requirements on minimum EC and/or delay-QoS exponent for their QoS-aware applications.

A. HOMOGENEOUS SYSTEM

In a homogeneous system, $\mu_u^{min} = \mu^{min}$ and $\theta_u = \theta, \forall u \in \mathcal{U}$. We start by solving the unconstrained optimization problem, i.e., the optimization problem (8) without the constraint (8b). From Proposition 1, at optimality of this problem, each subcarrier is given to the user with the best channel gain on that subcarrier, i.e., $\alpha_{v_n,n} = 1$, where $v_n = \text{argmax}_{u \in \mathcal{U}} \gamma_{u,n}, \forall n \in \mathcal{N}$. Let Ω_{v_n} be the set of subcarriers assigned to user v_n and $|\Omega_{v_n}|$ be the cardinality of the set $\Omega_{v_n}, \forall v_n \in \mathcal{U}$. Mathematically, $\Omega_{v_n} = \{n \in \mathcal{N}$ such that $\alpha_{v_n,n} = 1\}$. Once the subcarrier allocation is performed, the multi-user OFDMA system can be viewed as a single-user OFDM system. Next, we determine the transmit power allocated to each subcarrier so as to maximize the utility-pricing function of the system. Hence, problem (8) without the constraint on the minimum EC of each user, after the subcarrier allocation, can be rewritten as:

$$\max_{p\in\mathbb{R}_+^N} \eta^{av}(p) = \mathbb{E}\left\{\sum_{n=1}^{N} r_{v_n,n} - \beta p_{v_n,n}\right\}, \tag{16}$$

where $p \in \mathbb{R}_+^N$, is a vector comprising the allocated power to each subcarrier n, which is already assigned to user v_n. Problem (16) is concave in p. Hence, from (10), at optimality, we have

$$p_n^* = p_{v_n,n}^* = \left[\frac{1}{\beta \ln 2} - \frac{1}{\gamma_n}\right]^+, \quad \forall n \in \mathcal{N}. \tag{17}$$

Thus, the transmit power adaptation is performed using water-filling policy in frequency domain with the water-filling level equivalent to $\frac{1}{\beta \ln 2}$. Further, $r_{v_n,n}^* = \log_2(1 + \gamma_{v_n,n}p_{v_n,n}^*)$, $\forall n \in \mathcal{N}$ and $\eta^{av*} = \sum_{n=1}^{N}(r_{v_n,n}^* - \beta p_{v_n,n}^*)$. The obtained EC by this water-filling power allocation, $\mu_u^{wf} = -\frac{1}{\theta T_f B}\ln\left(\mathbb{E}\left\{e^{-\theta T_f b \sum_{m\in\Omega_u} r_{u,m}^*}\right\}\right)$, $\forall u \in \mathcal{U}$. For a homogeneous system with i.i.d channels, $\mu_u^{wf} = \mu^{wf}, \forall u \in \mathcal{U}$. If $\mu_u^{wf} = \mu^{wf} \geq \mu^{min}$, then the aforementioned solution, where subcarriers are assigned to users with maximum CNR and the power adaptation is performed with water-filling policy, is feasible to problem (8) and hence, optimal. However, this solution is no longer feasible when μ^{min} is set to a value greater than μ^{wf}.

If this global solution is no longer feasible then, at optimality, constraint (8b) is met with equality (from Proposition 2). Hence, using the Lagrangian approach, Lagrangian function, $\mathcal{L}$, of (8) can be written as:

$$\mathcal{L} = \mathbb{E}\left\{\sum_{u=1}^{U}\sum_{n=1}^{N}\alpha_{u,n}(r_{u,n} - \beta p_{u,n}) - \sum_{u=1}^{U}\lambda_u\left(e^{-\theta_u T_f b\sum_{m\in\mathcal{N}}\alpha_{u,m}r_{u,m}} - e^{-\theta_u T_f B\mu_u^{min}}\right)\right\}, \tag{18}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ is a vector of Lagrange multipliers $\lambda_u, \forall u \in \mathcal{U}$. For a homogeneous system, these multipliers are equal, i.e., $\lambda_u = \lambda, \forall u \in \mathcal{U}$ [19]. Thus, (18) for a homogeneous system can be re-written as:

$$\mathcal{L} = \mathbb{E}\left\{\sum_{u=1}^{U}\sum_{n=1}^{N}\alpha_{u,n}(r_{u,n} - \beta p_{u,n}) - \sum_{u=1}^{U}\lambda\left(e^{-\theta T_f b\sum_{m\in\mathcal{N}}\alpha_{u,m}r_{u,m}} - e^{-\theta T_f B\mu^{min}}\right)\right\}. \tag{19}$$

For some given power allocation, (19) is an increasing function of rates and accordingly, channel power gains. Hence, maximum value of $\mathcal{L}$ is obtained when subcarriers are assigned to users with best channel gains.

Let Ω_u be the set of channels assigned to user u, i.e., $\Omega_u = \{n \in \mathcal{N}$ such that $\alpha_{u,n} = 1\}$, $\theta_u' = \theta_u T_f b$ and $G_u = e^{-\theta_u T_f B\mu_u^{min}}$. For a homogeneous system, $G_u = G$ and $\theta_u' = \theta', \forall u \in \mathcal{U}$. Next, we determine the transmit power for user u over subcarriers $n \in \Omega_u$. We maximize the part of the Lagrangian $\mathcal{L}$ related to user u:

$$\mathcal{L}_u = \mathbb{E}\left\{\sum_{n\in\Omega_u}(r_{u,n} - \beta p_{u,n}) - \lambda(e^{-\theta'\sum_{m\in\Omega_u} r_{u,m}} - G)\right\} \tag{20}$$

and $\mathcal{L} = \sum_{u=1}^{U}\mathcal{L}_u$. Here, $\mathcal{L}_u$ is concave. Hence, the KKT conditions are both necessary and sufficient for optimality. At optimality, we have

$$\begin{aligned}\left.\frac{\delta\mathcal{L}_u}{\delta p_{u,n}}\right|_{P=P^*} &= 0, \quad \forall n \in \Omega_u,\\ \frac{\gamma_{u,n}}{\ln 2(1+\gamma_{u,n}p_{u,n}^*)} + \lambda e^{-\theta'\sum_{m\in\Omega_u} r_{u,m}^*}\frac{\theta'\gamma_{u,n}}{\ln 2(1+\gamma_{u,n}p_{u,n}^*)} &= \beta, \quad \forall n \in \Omega_u,\\ \frac{\gamma_{u,n}}{(1+\gamma_{u,n}p_{u,n}^*)} &= \frac{\beta\ln 2}{1+\theta'\lambda e^{-\theta'\sum_{m\in\Omega_u} r_{u,m}^*}}, \quad \forall n \in \Omega_u.\end{aligned} \tag{21}$$

Hence,

$$\frac{\gamma_{u,n}}{(1+\gamma_{u,n}p_{u,n}^*)} = \frac{\gamma_{u,n'}}{(1+\gamma_{u,n'}p_{u,n'}^*)}, \quad \forall n, n' \in \Omega_u. \tag{22}$$

Next, using (22), we have

$$\begin{aligned}e^{-\theta'\sum_{m\in\Omega_u} r_{u,m}^*} &= \prod_{m\in\Omega_u}(1+\gamma_{u,m}p_{u,m}^*)^{-\theta'/\ln 2},\\ &= \left(\prod_{m\in\Omega_u}\gamma_{u,m}^{-(\theta'/\ln 2)}\right) \times \left(\frac{1+\gamma_{u,n}p_{u,n}^*}{\gamma_{u,n}}\right)^{-|\Omega_u|\theta'/\ln 2}, \quad \forall n \in \Omega_u.\end{aligned} \tag{23}$$

Substituting (23) in (21), we get

$$\beta\ln 2\frac{(1+\gamma_{u,n}p_{u,n}^*)}{\gamma_{u,n}} = \left(1 + \theta'\lambda\left(\frac{1+\gamma_{u,n}p_{u,n}^*}{\gamma_{u,n}}\right)^{-|\Omega_u|\theta'/\ln 2} \times \prod_{m\in\Omega_u}\gamma_{u,m}^{-(\theta'/\ln 2)}\right), \quad \forall n \in \Omega_u. \tag{24}$$

Let $\zeta(\boldsymbol{\gamma}) = \gamma_{u,n}/(1+\gamma_{u,n}p_{u,n}^*)$, $K_u = \theta'\lambda\prod_{m\in\Omega_u}\gamma_{u,m}^{-(\theta'/\ln 2)}$, $Z = \beta\ln 2$ and $Y_u = |\Omega_u|\theta'/\ln 2$. Then, (24) can be written as:

$$\begin{aligned}\frac{Z}{\zeta(\boldsymbol{\gamma})} &= (1 + K_u\zeta(\boldsymbol{\gamma})^{Y_u}),\\ K_u\zeta(\boldsymbol{\gamma})^{Y_u+1} + \zeta(\boldsymbol{\gamma}) - Z &= 0.\end{aligned} \tag{25}$$

TABLE 1. Optimal subcarrier and power allocation in homogeneous systems.

Step	Algorithm
Step 1	Allocate subcarrier n to the user with the highest channel gain on that subcarrier, i.e., $\alpha_{v_n,n}=1$, where $v_n = \mathrm{argmax}_{u\in\mathcal{U}}\,\gamma_{u,n}$, $\forall n\in\mathcal{N}$.
Step 2	Compute the globally optimal power $p^*_{v_n,n}$, $\forall n \in \mathcal{N}$, using (10c).
Step 3	Check the feasibility of this globally optimal solution: If the globally optimal power is less than the power obtained from (26) then it is infeasible.
Step 4	In case of infeasibility, the feasible optimal power allocation is computed using (26).
Step 5	Update the Lagrange multiplier λ_u, $\forall u\in\mathcal{U}$, using (27) and repeat steps 1 - 5 for the next time slot.

Once $\zeta(\boldsymbol{\gamma})$ is evaluated, optimal value of power can be computed as:

$$p^*_n = p^*_{u,n} = \left[\frac{1}{\zeta(\boldsymbol{\gamma})} - \frac{1}{\gamma_{u,n}}\right]^+, \quad \forall n \in \Omega_u. \tag{26}$$

However, (25) is non-linear in $\zeta(\boldsymbol{\gamma})$ and can yield multiple solutions. Because of infeasibility, we ignore the values of $\zeta(\boldsymbol{\gamma})$ that yield negative power. Since the utility-pricing function is a decreasing function of power beyond global optimum (from Proposition 2), the optimal $\zeta(\boldsymbol{\gamma})$ is the one that gives minimum value of power. Hence, optimal feasible power will be the minimum non-negative power obtained from (26). Since the system is homogeneous, $\mathcal{L}_u$ will be same for all the users and hence, maximizing $\mathcal{L}_u$ is equivalent to maximizing $\mathcal{L}$. Further, stochastic dual gradient iterations used in [19], which is given as:

$$\lambda_u^{t+1} = \lambda_u^t + \epsilon^t\left(e^{-\theta'\sum_{m\in\Omega_u^t}\log_2(1+p^t_{u,m}\gamma^t_{u,m})} - e^{-\theta' N\mu^{min}}\right), \quad \forall u\in\mathcal{U} \tag{27}$$

can be used to compute the Lagrange multipliers after each time slot t and $\lim_{t\to\infty}\lambda_u^t = \lambda, u \in \mathcal{U}$. Here, Ω_u^t is the set of subcarriers in slot t for which user u has the largest channel gains and $\{\epsilon^t\}_{t=1,2,\ldots}$ is a positive learning sequence satisfying the following conditions:

$$\sum_{t=1}^{\infty}\epsilon^t = \infty; \quad \sum_{t=1}^{\infty}(\epsilon^t)^2 < \infty. \tag{28}$$

The convergence of these stochastic approximation iterations is shown in [19]. The iterative algorithm finding solution for optimal subcarrier and power allocation in homogeneous systems can be summarized in the Table 1.

B. HETEROGENEOUS SYSTEM

The users may have different requirements on minimum EC and/or delay-QoS exponent in a heterogeneous system. However, the channel gains, $h^t_{u,n}$, $\forall u, n, t$, are still i.i.d with the PDF f. It may be very difficult to derive the optimal subcarrier and power allocation algorithm for these systems due to heterogeneity in the QoS requirements of users. Hence, we propose two reduced-complexity sub-optimal algorithms for these systems.

1) ALGORITHM-I

Algorithm-I is an extension of the aforementioned optimal algorithm for homogeneous systems to heterogeneous systems. Here, each subcarrier n is allocated to the user with the best channel gain. Then, we solve the unconstrained optimization problem and compute the globally optimal power $p^*_{v_n,n}$ from (10c). If $p^*_{v_n,n}$ is feasible, we have an optimal solution. Else, we solve the Lagrangian function, $\mathcal{L}$, given by (18). The solution to this function is similar to the homogeneous system except for the fact that θ_u and μ_u can be different for different users. Thus, from (25) and (26), we have:

$$K_u\zeta(\boldsymbol{\gamma})^{Y_u+1} + \zeta(\boldsymbol{\gamma}) - Z = 0 \quad \text{and} \tag{29a}$$

$$p_n = p_{u,n} = \left[\frac{1}{\zeta(\boldsymbol{\gamma})} - \frac{1}{\gamma_{u,n}}\right]^+, \quad \forall n \in \Omega_u, \tag{29b}$$

where $\zeta(\boldsymbol{\gamma}) = \gamma_{u,n}/(1+\gamma_{u,n}p_{u,n})$, $Z = \beta\ln 2$, $K_u = \theta'_u\lambda_u\prod_{m\in\Omega_u}\gamma_{u,m}^{-(\theta'_u/\ln 2)}$ and $Y_u = |\Omega_u|\theta'_u/\ln 2$. Note that the subcarrier allocation to the user with the maximum CNR may not be optimal anymore. This is because, in a heterogeneous system, $\mathcal{L}$ does not necessarily increase with the channel power gain due to variation in θ_u and μ_u^{min}, $\forall u \in \mathcal{U}$.

Next, we perform the complexity analysis of Algorithm-I. For the subcarrier allocation, we need to locate the user from the set of $\mathcal{U}$ with the highest CNR in each subcarrier. Since there are N subcarriers, UN computations are required for subcarrier allocation. Next, we adapt the total transmit power for each subcarrier using (10c) and check its feasibility which needs the computation of $O(N)$. If infeasible, we evaluate the total power using (29b) which requires the complexity of $O(N)$. Further, updating Lagrange multiplier for each user mandates the complexity of $O(U)$. Hence, overall complexity of Algorithm-I is of $O(UN)$. Algorithm-I, though simple and optimal for a homogeneous system, fails to consider the heterogeneity in the QoS requirements of users and may not be suitable for a hetrogeneous system. Hence, we propose heterogeneity-aware Algorithm-II for heterogeneous systems.

2) ALGORITHM-II

In Algorithm-II, we first compute the globally optimal solution by allocating each subcarrier to the user with the best channel gain and then adapting the transmit power on each subcarrier using water-filling policy (10c). If this globally optimal solution is infeasible, we propose a new subcarrier allocation technique which takes into account the heterogeneity in the QoS requirements of users. Since the effect of QoS requirement is reflected in the required power and the quality of the link is reflected in the CNR, each user computes and conveys the required rate, which is an increasing function of required power and CNR, to the BS. To compute the required rate, each user in a scheduling interval/slot assumes that it has all the subcarriers allocated to it, i.e., $\Omega_u = \mathcal{N}, \forall u \in \mathcal{U}$. With this assumption, each user u computes the power required on subcarrier n, $p_{u,n}^{req}, \forall u \in \mathcal{U}, n \in \mathcal{N}$, to meet its EC using (29b). The required rate of user u on subcarrier n on that scheduling interval is then given by:

$$r_{u,n}^{req} = \log_2(1 + \gamma_{u,n} p_{u,n}^{req}), \quad \forall u \in \mathcal{U}, \ n \in \mathcal{N}. \tag{30}$$

After the computation of required rate, each user conveys this information to the BS. The BS assigns a subcarrier n to the user v_n which has the highest required rate on that subcarrier, i.e., $\alpha_{v_n,n} = 1$ and $\alpha_{u,n} = 0, \forall u \in \mathcal{U} \backslash \{v_n\}$, where $v_n = \text{argmax}_{u \in \mathcal{U}} r_{u,n}^{req}$. The required rate of user u on subcarrier n is high if it has higher CNR ($\gamma_{u,n}$) on that subcarrier or it requires higher power ($p_{u,n}^{req}$). Requirement of higher power indicates that the user is currently far from meeting its QoS requirement and hence, has a higher chance of getting the subcarrier. Therefore, assigning the subcarrier n to the user with highest $r_{u,n}^{req}$ results in power savings due to better channel condition and/or QoS requirement satisfaction. Once the subcarrier allocation is performed, power adaptation policy (29b) can be used to compute the total power.

Further, we evaluate the computational complexity of Algorithm-II. UN computations are required for subcarrier allocation based on CNR and complexity of $O(N)$ is needed to adapt the globally optimal transmit power and check its feasibility. If infeasible, $2UN$ computations are required to evaluate $p_{u,n}^{req}$ and $r_{u,n}^{req}, \forall u \in \mathcal{U}, n \in \mathcal{N}$. Conveying this information to the BS accounts for additional uplink (UL) data. For subcarrier allocation, BS needs to find the highest required rate for each subcarrier over all the users, which mandates UN more computations. Then, we adapt the total power for each subcarrier, which requires $O(N)$ computations. Overall, the complexity of Algorithm-II is $O(4UN)$ with some information overhead.

An exhaustive search has to be performed to compute the optimal solution in a heterogeneous system if the global optimum is infeasible. Since it is assumed that a subcarrier is assigned to only one user in a scheduling interval, there are U^N possible subcarrier assignments. For a given subcarrier assignment, (29b) can be used for power distribution among allocated user-subcarrier pairs which need the computations of $O(N)$. The optimal value is the maximum value of utility-pricing function computed over all U^N possible subcarrier assignment schemes and the corresponding subcarrier and power allocation is the optimal scheduling scheme for the system. Hence, the complexity of the optimal algorithm in a heterogeneous system is of $O(NU^N)$.

IV. ILLUSTRATIVE RESULTS

The simulation scenario consists of a single cell with radius 500 m. BS is located at the center of the cell and users are uniformly distributed in the cell. We simulate a unit-variance Rayleigh channel for each user. Hence, channel gains $h_{u,n}, \forall u \in \mathcal{U}, n \in \mathcal{N}$ are exponentially distributed. We use the results from Section III to optimize the ergodic utility-pricing function of the system subject to constraints on the minimum EC of each user in the system. The bandwidth of the subcarrier, b, is considered to be 200 KHz and the system bandwidth $B = 200N$ KHz. Further, the noise spectral density, N_0, is -174 dBm/Hz and we run the simulation for 20, 000 time slots. Also, the distance-based path loss [29] for a macro cell environment considering a carrier frequency of 2 GHz, in dB, is given by:

$$Pl_u^{dB} = 128.1 + 37.6 \log_{10}(d_u), \tag{31}$$

where d_u is the distance of the user from the BS in km. Hence, $Pl_u = 10^{(Pl_u^{dB}/10)}$. The CPU processor used during simulation is a core duo processor with a speed of 1.83 GHz and a memory of 1 GB RAM, the software used for simulation is MATLAB 8.2, and the precision level of bisection algorithm is 10^{-6}.

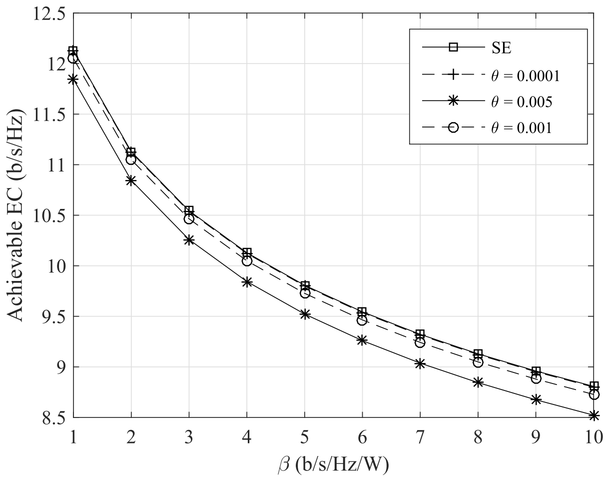

FIGURE 2. Achievable EC (μ^{wf}) using water-filling power allocation versus β for different θ when $N = 64$, $U = 1$ and CNR = 35 dB.

First, we perform simulations to observe the achievable EC using water-filling power allocation (μ^{wf}), which gives maximum achievable SE in the system, as a function of β and θ. We can see in Fig. 2 that μ^{wf} decreases with the increase in β. This is because higher values of β correspond to higher penalties in power which result in less allocated power and hence, smaller achievable EC. Further, we investigate the possible values of μ^{wf} for different θ and observe that achievable EC decreases with the increase in θ. For example, when $\beta = 8$ b/s/Hz/W, μ^{wf} decreases from 9.12 b/s/Hz to 8.84 b/s/Hz as θ is increased from 0.0001 to 0.001. If $\mu^{min} \leq \mu^{wf}$ for these values of θ and β, water-filling power allocation in (10c) is feasible and hence, the optimal solution. Else, the solution is given by the power allocation (26). Note that when $\theta \rightarrow 0$, EC of the system approaches SE.

Next, we perform simulations to investigate the behaviour of the power or signal-to-noise ratio (SNR) as a function of minimum required EC (μ^{min}) and β when $U = 1, N = 64$ and the average CNR = 35 dB. We define SNR as the

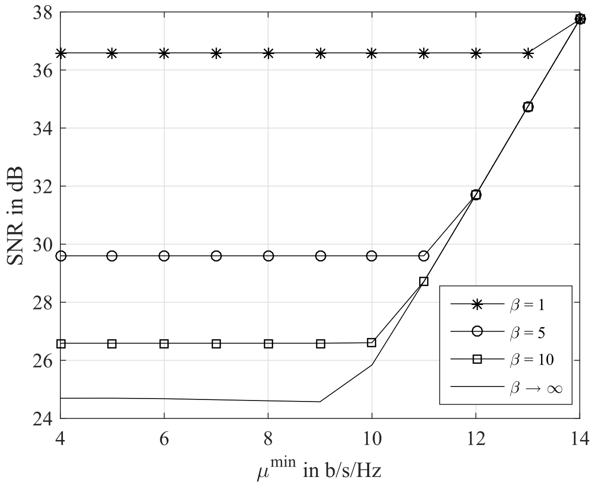

FIGURE 3. SNR versus μ^{min} for different β when $\theta = 0.0001$ and $N = 64$, $U = 1$ and CNR = 35 dB.

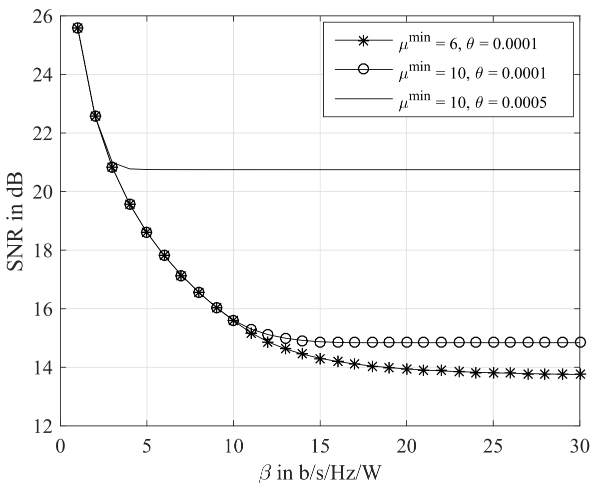

FIGURE 4. SNR versus β for different μ^{min} and θ when $N = 64$, $U = 1$ and CNR = 24 dB.

product of power and CNR in linear scale or accordingly, sum of power and CNR in logarithmic scale. We can observe from Fig. 3 that the SNR (or power) is invariant to the imposed constraint on minimum EC for its lower values. This is because, for lower μ^{min}, water-filling power allocation in (10c) obtained by solving the unconstrained problem satisfies the constraint (8b) and hence, optimal. With further increment in μ^{min}, this water-filling power allocation is no longer feasible and the power is given by (26). We can see in Fig. 3 that the power or SNR (in logarithmic scale) is then linearly increased with μ^{min}.

Moreover, in Figs. 3 and 4, we can see the impact of β on the power (or SNR). The higher values of β correspond to higher penalties in power. The power (or SNR) is a decreasing function of β because of higher penalty. When β is set to higher values, maximization of the utility-pricing function is equivalent to minimization of the transmit power. Thus, for the systems with higher μ^{min} or higher β, only QoS-aware traffic can be supported in the system.

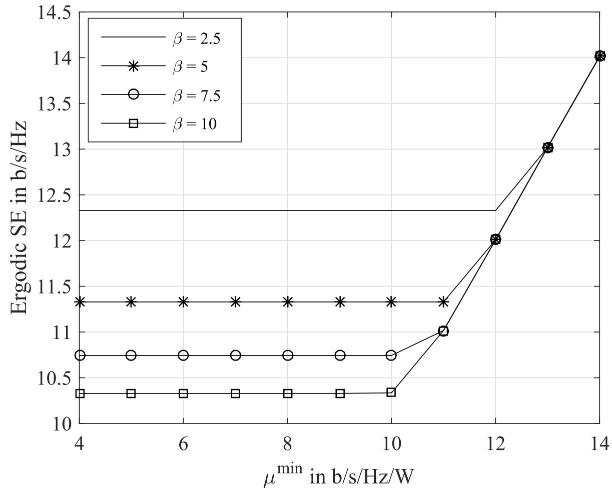

FIGURE 5. Ergodic spectral efficiency versus μ^{min} for different values of β when $\theta = 0.0001$, $N = 64$, $U = 1$ and CNR = 35 dB.

In Fig. 5, we investigate the impact of μ^{min} and β on the SE of the system. As aforementioned, the power for the lower values of μ^{min} is given by (10c) and hence, is indifferent to variation in μ^{min}, but decreases with the increase in β. Thus, SE as a function of μ_{min}, for a given β, is constant for the lower values of μ_{min}. Further increment in μ^{min} leads to linear increment in the required SE. SE of the system is a decreasing function of β, but only for the lower values of μ^{min} when BE traffic can be supported in the system. Moreover, Fig. 5 gives an insight on the amount of BE traffic that can be supported in the system. For example, when $\mu^{min} = 6$ b/s/Hz and $\beta = 5$ b/s/Hz/W, SE of the system is 11.40 b/s/Hz. Hence, the achievable SE for the BE traffic is 5.40 b/s/Hz. When $\beta = 5$ b/s/Hz/W and $\mu^{min} = 12$ b/s/Hz, SE of the system is also 12 b/s/Hz and hence, BE traffic cannot be supported.

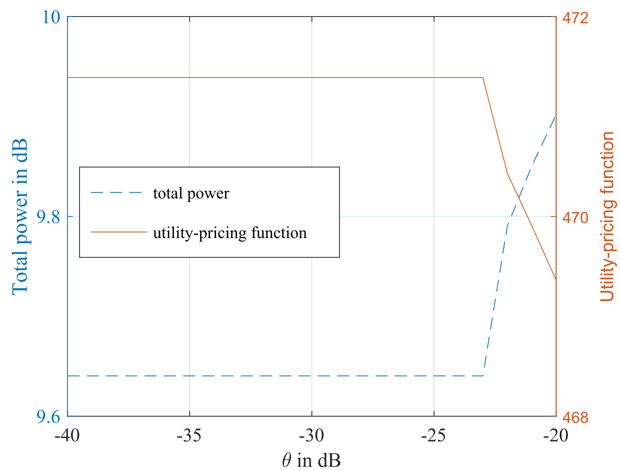

FIGURE 6. Power and utility-pricing function versus θ when $\beta = 10$ b/s/Hz/W, $\mu^{min} = 8$ b/s/Hz, $U = 1$, $N = 64$ and CNR = 35 dB.

Furthermore, we plot the variation in the total power and the utility-pricing function as a function of θ for a given μ^{min} and β when $U = 1$ and $N = 64$. We can see in Fig. 6 that, for small θ, power obtained from (10c) is feasible and hence,

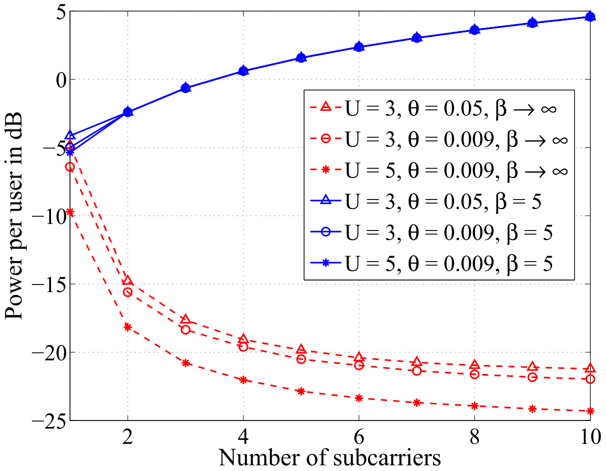

FIGURE 7. **Power per user versus number of subcarriers in a homogeneous system for various U, θ and β, in b/s/Hz/W, when CNR = 24 dB and μ^{min} = 8 b/s/Hz for each user on each subcarrier.**

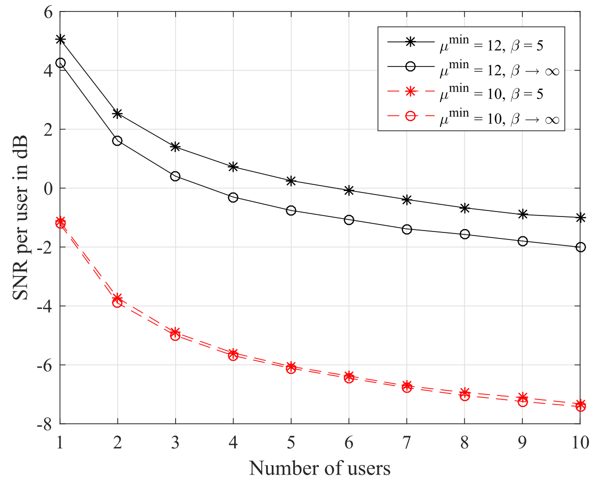

FIGURE 8. **SNR per user versus number of users in a homogeneous system for various μ_{min}, in b/s/Hz, and β, in b/s/Hz/W, when $N = 64$, CNR = 35 dB and $\theta = 0.02$ for each user on each subcarrier.**

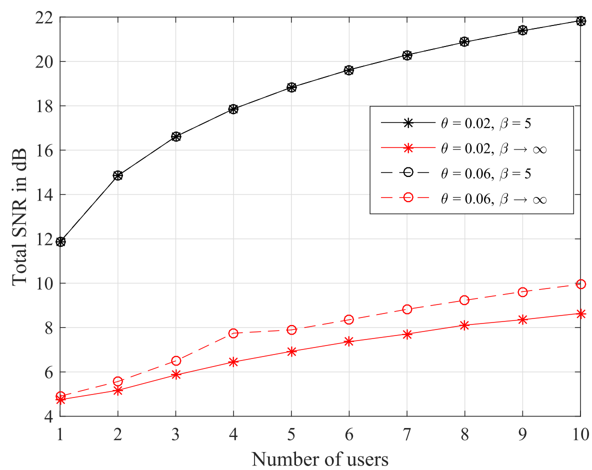

FIGURE 9. **Total SNR versus number of users in a homogeneous system for various θ and β, in b/s/Hz/W, when $N = 64$, CNR = 35 dB and μ^{min} = 12 b/s/Hz for each user on each subcarrier.**

the total power and the utility-pricing function are indifferent to variation in θ. As θ increases, the system becomes more stringent and hence, $\mu^{wf} < \mu^{min}$. The total power is then given by (26) which increases with the increase in θ while the utility-pricing function decreases with the increase in θ.

In Fig. 7, we can observe the variation in power per user, in dB, as a function of N for various U, θ and β when $\mu^{min} =$ 8 b/s/Hz and average CNR = 24 dB on each subcarrier for each user in a homogeneous system, i.e., $\mu_u^{min} = \mu^{min}$, $\theta_u = \theta, \forall u \in \mathcal{U}$. Since the power for each user is a vector in $\mathbb{R}_+^N$, we plot the sum of this power vector over subcarriers. As already mentioned, when $\beta \to \infty$, the problem of utility-pricing maximization is equivalent to power minimization. For these systems, minimum power required by each user, in dB, to meet its EC is an exponentially decreasing function of N. This is because of the increment in the system bandwidth and frequency diversity. Frequency diversity is obtained because we assume that the coherence bandwidth of the system is of the order of subcarrier bandwidth b. Moreover, for smaller β, the globally optimal power given by (10c) is feasible which is water-filling in nature. We observe that the power per user for these kind of systems increases with the increase in N due to frequency diversity and both the QoS-aware and BE traffic can be supported in the system.

Further, in Figs. 8 and 9, we plot variation in SNR and accordingly, power as a function of U for different μ_{min}, θ and β in a homogeneous system when $N = 64$ and average CNR = 35 dB for each user on each subcarrier. The SNR values in Fig. 8 represent the SNR required by each user. We observe that, for systems with higher values of β, the minimum SNR required by each user to meet its EC constraint decreases with the increase in number of users in the system. This is because of multi-user diversity. When the number of UEs increases, BS, in each scheduling interval, has a larger pool of UEs to choose from and thus, the probability of finding a UE with good channel condition increases, reducing the minimum required SNR (or power) per user. However, as the number of users increases, BS also has more QoS-aware traffic in the system whose QoS has to be satisfied. We can observe in Fig. 9 that the SNR of the BS (sum of SNR for each user) increases almost linearly with the number of users in the system. Moreover, for systems with smaller β, the impact of multi-user diversity on the SNR per user is insignificant.

Next, we examine the effectiveness of the proposed sub-optimal algorithms for heterogeneous systems in comparison with the optimal algorithm. Optimal utility-pricing function is obtained by the exhaustive search over all U^N possible subcarrier assignments and has a complexity of order $O(NU^N)$. Table 2 shows that the Algorithm-I and Algorithm-II provide near-optimal utility-pricing function with a low-computational complexity of $O(UN)$ and $O(4UN)$ with some information overhead, respectively. $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{min}$ in Table 2 are vectors in $\mathbb{R}_+^U$ representing the delay-QoS

TABLE 2. **Comparison of utility-pricing function of sub-optimal algorithms with optimal algorithm for various values of N and U in a heterogeneous system when $\beta = 10$.**

U	N	$\boldsymbol{\theta}$	$\boldsymbol{\mu}_{min}$ in b/s/Hz	Utility-pricing function in b/s/Hz for Algorithm-I	Algorithm-II	Optimal Algorithm
2	2	[0.02, 0.09]	[8,13]	14.8556	16.6499	17.4575
2	3	[0.02, 0.09]	[8,13]	21.0808	24.0146	24.7126
3	2	[0.02, 0.005 ,0.09]	[8, 15 ,13]	12.7748	15.6932	17.7815
3	3	[0.02, 0.005 ,0.09]	[8, 15 ,13]	23.0933	23.6693	24.9882
4	2	[0.01, 0.09, 0.03, 0.06]	[8,7,10, 5]	17.4742	18.3906	19.2091
4	3	[0.01, 0.09, 0.03, 0.06]	[8,7,10, 5]	24.6749	25.7512	26.8288

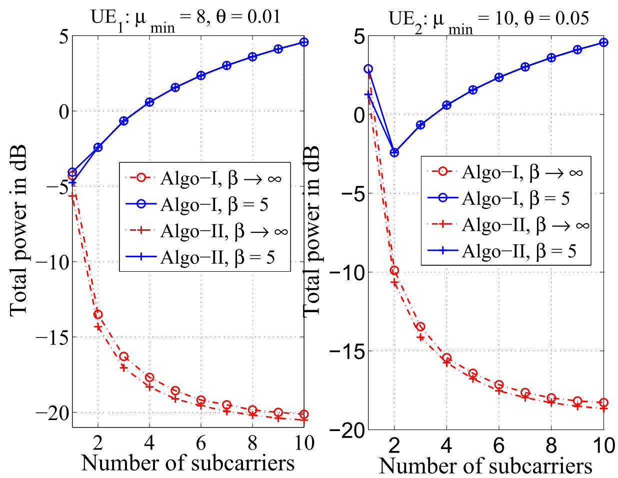

FIGURE 10. **Total power versus number of subcarriers in a heterogeneous system for Algorithm-I and Algorithm-II when $U = 2$ and CNR = 24 dB for each user on each subcarrier.**

exponent and minimum EC of each user, respectively. Moreover, in Fig. 10, we observe the effectiveness of Algorithm-II over Algorithm-I. We consider two users, UE_1 and UE_2, with $\boldsymbol{\mu}_{min} = [8, 10]$ and $\boldsymbol{\theta} = [0.01, 0.05]$. We observe that, for higher β or smaller N, Algorithm-II outperforms Algorithm-I since the resource allocation in Algorithm-II, unlike Algorithm-I, takes into account the heterogeneity in QoS requirements along with the channel conditions. However, this performance enhancement is obtained at the cost of additional complexity and information overhead. Furthermore, for smaller β and higher N, the globally optimal solution given by (10c) is likely to be feasible and hence, both the algorithms behave in a similar way.

V. CONCLUSION

In this paper, we formulated an optimization problem to maximize the utility-pricing function of a multi-user OFDM system where each user has the data corresponding to QoS-aware and BE applications subject to constraints on the minimum EC of QoS-aware traffic for each user. We considered cases with homogeneity and heterogeneity in the system and proposed optimal and sub-optimal algorithms for resource allocation, respectively. We observed that the reduced-complexity algorithms for heterogeneous systems provide near-optimal solutions with much less complexity. To add, we illustrated the effectiveness of Algorithm-II over Algorithm-I for the systems with higher β or μ_{min} where the globally optimal solution is not feasible. Further, we observed that the transmit power is a decreasing function of β for its smaller values while is invariant to μ^{min} for its lower values. Moreover, transmit power increases exponentially with the increase in μ^{min} and is invariant to β for its higher values. For higher μ^{min} and β, BE traffic cannot be supported in the system. We also observed that the transmit power for the systems with higher β decreases, while increases for the system with smaller β, as a function of N due to increment in the system bandwidth and the frequency diversity in the system.

REFERENCES

[1] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
[2] C. S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.
[3] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.
[4] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.
[5] J. Yuan and Q. Wang, "Delay quality-of-service driven resource allocation for relay-based multiuser OFDMA networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2012, pp. 3889–3894.
[6] K. Pentikousis, "In search of energy-efficient mobile networking," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 95–103, Jan. 2010.
[7] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.
[8] C. Isheden and G. P. Fettweis, "Energy-efficient link adaptation on parallel channels," in *Proc. 19th Eur. Signal Process. Conf.*, Sep. 2011, pp. 874–878.
[9] G. Miao, N. Himayat, and G. Y. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 545–554, Feb. 2010.
[10] G. Miao, N. Himayat, Y. Li, and D. Bormann, "Energy-efficient design in wireless OFDMA," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 3307–3312.
[11] S. Khakurel, L. Musavian, and T. Le-Ngoc, "Energy-efficient resource and power allocation for uplink multi-user OFDM systems," in *Proc. IEEE PIMRC*, Sep. 2012, pp. 357–361.
[12] F. Haider, C.-X. Wang, H. Haas, E. Hepsaydir, and X. Ge, "Energy-efficient subcarrier-and-bit allocation in multi-user OFDMA systems," in *Proc. IEEE VTC Spring*, Yokohama, Japan, May 2012, pp. 1–5.
[13] S. Zhang, Q. Wu, S. Xu, and G. Y. Li, "Fundamental green tradeoffs: Progresses, challenges, and impacts on 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 33–56, 1st Quat., 2017.

[14] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.
[15] L. Musavian and T. Le-Ngoc, "Energy-efficient power allocation over Nakagami-*m* fading channels under delay-outage constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4081–4091, Aug. 2014.
[16] M. C. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channels under QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4252–4263, Aug. 2009.
[17] J. Chen and K.-K. Wong, "Improving energy efficiency for multiuser MIMO systems with effective capacity constraints," in *Proc. IEEE Veh. Technol. Conf.*, Apr. 2009, pp. 1–5.
[18] Q. Wu, W. Chen, M. Tao, J. Li, H. Tang, and J. Wu, "Resource allocation for joint transmitter and receiver energy efficiency maximization in downlink OFDMA systems," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 416–430, Feb. 2015.
[19] K. T. Pan, T. Le-Ngoc, and L. B. Le, "Optimal resource allocation for buffer-aided relaying with statistical QoS constraint," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 959–972, Mar. 2016.
[20] L. Chen *et al.*, "Utility-based resource allocation for mixed traffic in wireless networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Sep. 2011, pp. 91–96.
[21] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 66–71, Jan. 2009.
[22] H. Holma and A. Toskala, *LTE for UMTS: OFDMA and SC-FDMA Based Radio Access*. Hoboken, NJ, USA: Wiley, 2009.
[23] S. Khakurel, T. Le-Ngoc, and L. Musavian, "Qos-driven energy-efficient power adaptation in a multi-channel fading communication link," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 4038–4043.
[24] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[25] J. A. Bucklew, *Introduction to Rare Event Simulation*. New York, NY, USA: Springer-Verlag, 2004.
[26] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127–134, Dec. 2005.
[27] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utility-based power-control scheme in wireless cellular systems," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 210–221, Apr. 2003.
[28] T. Ibaraki and N. Katoh, *Resource Allocation Problems: Algorithmic Approaches*. Cambridge, MA, USA: MIT Press, Mar. 1988.
[29] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) System Scenarios Release 8*, document TR 36.942 V1.12.0, 3GPP, 2007.

SUMAN KHAKUREL received the B. Tech. degree (Hons.) in electronics engineering from Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India, in 2010, and the M.Eng. degree in electrical engineering from McGill University, Montreal, QC, Canada, in 2013.

From 2010 to 2011, he was a Junior Research Fellow with the Tata Teleservices IITB Centre of Excellence in Telecommunications, IIT Bombay, Bombay, India, where he was involved in research and development of cellular system evolution, as part of the Long-Term Evolution Team. He is currently a Senior Software Engineer with Thomson Reuters. His research interests include digital communications and dynamic resource management for wireless networks, with emphasis on heterogeneous networks, quality-of-service provisioning, and energy efficiency.

Mr. Khakurel received the 2012 Graduate Excellence Fellowship from McGill University and gold medals from SVNIT for his academic excellence.

LEILA MUSAVIAN (S'05–M'07) received the Ph.D. degree in telecommunications from King's College London, U.K., in 2006.

She was a Post-Doctoral Fellow with INRS-EMT, Canada, from 2006 to 2008, a Research Associate with Loughborough University, U.K., from 2009 to 2010, and a Research Associate with McGill University from 2011 to 2012. She was a Lecturer with the InfoLab21, Lancaster University, from 2012 to 2016. She is currently a Reader in telecommunications with the School of Computer Science and Electrical Engineering, University of Essex. Her research interests include radio resource management, low latency communications, next generation wireless networks, energy harvesting, green communication, energy-efficient transmission techniques, cross-layer design for delay QoS provisioning, and 5G systems. She has been a TPC member of several conferences, including the IEEE ICC, the IEEE GLOBECOM, the IEEE WCNC, the IEEE ICCCN, the IEEE PIMRC, and ChinaCom. She was the TPC Co-Chair of CorNer 2016 (in conjunction with ISWCS 2016) and the Co-Chair of mmWave 5G (STEMCOM 2016). She is an Editor of the IEEE Transactions on Wireless Communications, an Executive Editor of the *Transactions on Emerging Telecommunications Technologies*, and an Associate Editor of *Internet Technology Letters* (Wiley).

HUNG V. VU (S'14) received the B.Eng. degree (Hons.) in electronics and telecommunications from the Hanoi University of Technology, Hanoi, Vietnam, in 2011, and the M.Sc. degree (Hons.) in electrical and computer engineering from The University of Akron, Akron, OH, USA, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. His research interest lies in the areas of information theory and wireless communications.

THO LE-NGOC (F'97) received the B.Eng. degree in electrical engineering and the M.Eng. degree in 1976 and 1978, respectively, and the Ph.D. degree in digital communications from the University of Ottawa, Ottawa, ON, Canada, in 1983.

From 1977 to 1982, he was with Spar Aerospace Ltd., Sainte-Anne-de-Bellevue, QC, Canada, where he was involved in the development and design of satellite communications systems. From 1982 to 1985, he was an Engineering Manager of the Radio Group, Department of Development Engineering, SR Telecom Inc., Saint-Laurent, QC, Canada, where he developed the new point-to-multipoint demand-assigned time-division multiple-access/time-division multiplexing subscriber radio system SR500. From 1985 to 2000, he was a Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal. Since 2000, he has been with the Department of Electrical and Computer Engineering, McGill University. His research interests include broadband digital communications. He is a fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada. He received the 2004 Canadian Award in Telecommunications Research and the 2005 IEEE Canada Fessenden Award. He holds a Canada Research Chair (Tier I) on broadband access communications.