Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2010, Article ID 168357, 10 pages
doi:10.1155/2010/168357

Research Article

Uplink Cross-Layer Scheduling with Differential QoS Requirements in OFDMA Systems

Bo Bai,[1,2] Wei Chen,[2] Zhigang Cao,[2] and Khaled Ben Letaief[1]

[1] *Department of Electronic and Computer Engineering, The Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong*
[2] *Department of Electronic Engineering, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Bo Bai, eebob@ust.hk

Received 15 January 2010; Revised 29 June 2010; Accepted 21 September 2010

Academic Editor: Mohammad Shikh-Bahaei

Fair and efficient scheduling is a key issue in cross-layer design for wireless communication systems, such as 3GPP LTE and WiMAX. However, few works have considered the multiaccess of the traffic with differential QoS requirements in wireless systems. In this paper, we will consider an OFDMA-based wireless system with four types of traffic associated with differential QoS requirements, namely, minimum reserved rate, maximum sustainable rate, maximum latency, and tolerant jitter. Given these QoS requirements, the traffic scheduling will be formulated into a cross-layer optimization problem, which is convex fortunately. By separating the power allocation through the waterfilling algorithm in each user, this problem will further reduce to a kind of continuous quadratic knapsack problem in the base station which yields low complexity. It is then demonstrated that the proposed cross-layer method cannot only guarantee the application layer QoS requirements, but also minimizes the integrated residual workload in the MAC layer. To further enhance the ability of QoS assurance in heavily loaded scenario, a call admission control scheme will also be proposed. The simulation results show that the QoS requirements for the four types of traffic are guaranteed effectively by the proposed algorithms.

1. Introduction

Orthogonal frequency-division multiple access (OFDMA) offers a very attractive solution in providing high performance and flexible deployment for broadband wireless access network. In particular, OFDMA provides at more degrees of freedom for multiuser systems. The subcarriers can be allocated dynamically at different time instances to exploit the multiuser diversity [1] and frequency diversity [2], and adaptive power allocation can also be applied to further improve the power efficiency [3]. To enhance the efficiency and fairness, OFDMA also allows us to schedule time-domain resources, referred to as timeslots.

The typical OFDMA systems in wireless communications are 3GPP LTE-based cellular system [4] and IEEE 802.16 protocol-based WiMAX system [5]. These newly emerging systems provide a platform for applying the cross-layer resource allocation and scheduling technology. These systems are designed as a unified wireless access system to support multiple types of traffic, such as voice, data, audio/video, multimedia, interactive game, and Internet access. Thus, how to jointly use these technologies in the physical (PHY) layer and MAC layer to support the traffic with differential QoS requirements in the application layer is a central problem in OFDMA systems [6]. In this paper, we shall focus on this problem and use a cross-layer optimization methodology to provide a traffic scheduling method for supporting efficiently multiplexing services with a variety of QoS requirements.

Due to the stochastic nature of the traffic arrival process and the wireless channel, it is a challenging work to achieve fair and efficient resource allocation and QoS-guaranteed scheduling in OFDMA systems. In 1995, a joint-layer optimization perspective was proposed by Telatar and Gallager in [7]. Subsequently, Berry and Yeh put forward that the future wireless communication system design needs cross-layer optimization methodology [8]. They also discussed

the cross-layer approach for wireless resource allocation in multiaccess and broadcasting queueing systems, respectively. Specifically, in order to collect all the parameters together in the uplinks, one may formulate the system as a multiaccess queueing system or generic switch model and consider the weighted sum of the queue lengths, which is often referred to as the integrated workload. More recently, Stolyar proved the optimality of the MaxWeight scheduling in [9]. In [10], Mandelbaum and Stolyar extended this method to the continuous strictly increasing convex function of the queue length and proved the optimality of $C - \mu$ law scheduling. Based on the queueing theory and optimization method, Niyato and Hossain studied the radio resource management in IEEE 802.16 wireless broadband system [11]. An alternative method to incorporate concerns and constraints of various layers is to apply utility maximization formulation. In [12], Song et al. used this method to obtain a queue-aware and channel-aware scheduling algorithm, that is, transmit the traffic which minimizes the average delay. Based on the similar framework, Kulkarni and Rosenberg studied the opportunistic scheduling framework of multiple QoS requirements and short-term fairness in the system with multiple wireless interfaces [13]. In [14], Fu et al. solved the dual problems of maximizing expected throughput given limited energy and of minimizing expected energy given the minimum throughput constraint.

The above works have significantly enhanced the overall performance of wireless communications. However, they did not consider the scheduling problem of multiple types of traffic with differential QoS requirements, which is a practical scenario in OFDMA wireless access network. A typical OFDMA system, say IEEE 802.16 broadband wireless access network, has multiple independent users communicating with one base station (BS). There are four types of traffic in IEEE 802.16 protocol, namely, best effort service (BE), nonrealtime polling service (nrtPS), realtime polling service (rtPS), and unsolicited grant service (UGS) [5]. Any application-layer traffic must be classified into one of these types, and its QoS requirements can be described differentially by minimum reserved rate, maximum sustainable rate, maximum latency, and tolerant jitter. Thus, the arrival traffic of each user will be stored in different buffers and scheduled by a cross-layer scheduler in BS. Since the OFDMA-based PHY layer is timeslotted, every user should offer the traffic transmission request and its QoS parameters at the beginning of each timeslot. Given the constraints of QoS requirements and the instantaneous channel conditions, the scheduler allocates subcarriers, power, and timeslots, so as to transmit the traffic efficiently and guarantee the differential QoS requirements.

In this paper, the integrated residual workload method is introduced to cover the above considerations. By using this method, the resource allocation and traffic scheduling can be formulated into a cross-layer optimization problem under the transmission rate constraints, which is convex fortunately. Since the power allocation gives little advantage in terms of ergodic capacity [15], we decompose the power allocation from the original convex optimization problem through the water-filling algorithm in each user. The resulting optimization problem in BS, referred to as the time-frequency allocation problem, is fortunately a continuous quadratic knapsack problem with a generalized upper bound and an angular structure in the constraints. The knapsack problem (integer or continuous) has been studied for decades, which has often used to solve resource allocation problems in operational research, economics, military, and communications [16, 17]. According to the results in [18, 19], this time-frequency allocation problem can be solved with a low complexity. At this context, an integrated residual workload minimization (IRWM) algorithm and a heuristic call admission control (CAC) algorithm are proposed as a framework of the resource management scheme for future OFDMA-based wireless access networks. It is then demonstrated that the proposed cross-layer method cannot only guarantee the application layer QoS requirements, but also minimize the integrated residual workload in the MAC layer. The simulation results also verified that the QoS requirements for the four types of traffic are guaranteed effectively by the proposed scheduling algorithms.

The rest of the paper is organized as follows. Section 2 presents the system model and the QoS requirements. In Section 3, we present the cross-layer optimization problem and the problem decomposition. An optimal scheduling policy and a heuristic CAC algorithm is also presented in this section. Simulation results are presented in Section 4. Section 5 concludes this paper.

2. Cross-Layer Multiaccess Queuing Model

Consider an OFDMA system with multiple independent access users, where each user transmits four types of traffic to a BS. Then, each user has four queues, each of which corresponds to one type of traffic. In this system, each subcarrier can serve any queue, and each queue can be served by any subcarrier. Thus, the queues depend on each other and the subcarriers cannot be scheduled separately. Then, the uplink scheduling issue in this OFDMA system can be seen as a centralized cross-layer multiaccess queuing system, shown in Figure 1, which is also referred to as the generic switch model in [9].

2.1. QoS Parameters and Traffic Scheduling Framework. Similar to IEEE 802.16e protocol [5], the traffic supported by this OFDMA system is divided into four types, and a different traffic type has different QoS requirements. The QoS requirements supported include:

(i) minimum reserved rate (Min R), denoted by $R^{\min}$, which is the transmission rate that cannot be violated even the system is in congestion;

(ii) maximum sustainable rate (Max R), denoted by $R^{\max}$, which is the peak transmission rate allowed;

(iii) maximum latency (Max L), denoted by L, which is the maximum sojourn time of the traffic in a queue;

(iv) tolerant jitter (Tol J), denoted by J, which is the maximum absolute value of the latency difference for the same type of traffic.

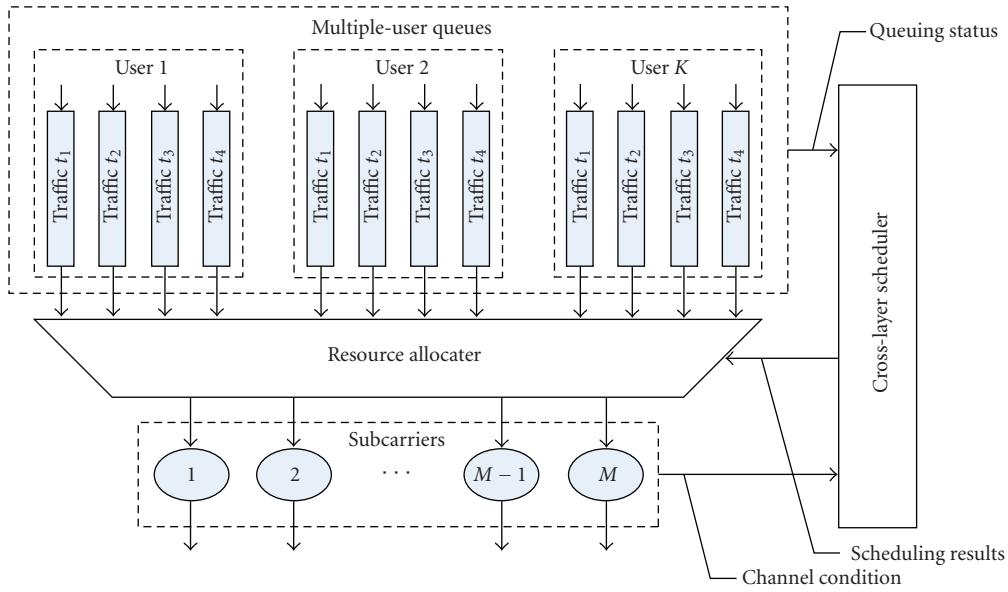

Figure 1: Cross-layer multiaccess queuing system for OFDMA systems.

We use $\mathcal{T}$, to denote the set of traffic types (in this paper, the script symbol $\mathcal{X}$ is used to denote a set, whose cardinality will be denoted by X), Then, the best effort (BE) service, denoted by $t_1 \in \mathcal{T}$, is used to support the best effort traffic, such as E-mail and file transfer. There are no explicit QoS requirements. The nonrealtime polling service (nrtPS), denoted by $t_2 \in \mathcal{T}$, assures the uplink service flow receives transmission opportunities even during network congestion, such as Internet browsing and data transfer. The QoS requirements supported include Min R and Max R. The realtime polling service (rtPS), denoted by $t_3 \in \mathcal{T}$, offers realtime uplink service flows that transport variable-size data packets, such as moving pictures experts group (MPEG) video, interactive game. The QoS requirements supported include Min R, Max R, and Max L. The unsolicited grant service (UGS), denoted by $t_4 \in \mathcal{T}$, offers realtime service flows that transport fixed-size data packets arriving periodically, such as T1/E1 and voice over IP without silence suppression. The QoS requirements supported include Min R, Max R (which is equal to Min R), Max L, and Tol J.

In the interested OFDMA system, access user must negotiate the QoS requirements with BS before the traffic connection is established. The negotiation process determines the value of $R^{\min}$, $R^{\max}$, L, and J for each type of traffic. Since this OFDMA system is timeslotted, then each user must provide the current value of the QoS parameters (including rate, latency, and jitter) and the traffic transmission request for each type of traffic at the beginning of every timeslot. Then, under the constraints of the QoS requirements and the channel conditions, BS determines which type and how much the traffic will be transmitted in this timeslot and allocates subcarrier, power, and time to them. Thus, the scheduling policy of BS is the central problem here. The cross-layer method proposed in the paper is an optimal resource allocation and scheduling method.

2.2. Problem Formulation. In the OFDMA system, we assume BS has the perfect channel sate information (CSI), since it can be achieved through ranging, channel estimation, and the message interaction between BS and users [5]. According to [20], the instantaneous capacity of subcarrier m for user k with adaptive modulation coding (AMC) mechanism is given by

$$C_{km} = B\log_2(1 + Q\gamma_{km}), \quad k \in \mathcal{K},\ m \in \mathcal{M}, \tag{1}$$

where B is the bandwidth of the subcarrier, $\mathcal{K}$ is the set of access users, and $\mathcal{M}$ is the set of subcarriers. The parameter Q is calculated by

$$Q = \frac{1.5}{-\ln(5\mathrm{BER})}, \tag{2}$$

where BER is the target bit error rate of the AMC mechanism. The instantaneous signal-to-noise ratio (SNR) γ_{km} can be rewritten as

$$\gamma_{km} = \beta_{km}|h_{km}|^2\mathrm{SNR}_k, \quad k \in \mathcal{K},\ m \in \mathcal{M}, \tag{3}$$

where SNR_k is the average SNR of the receiver in user k, β_{km} is the proportion of the power allocated to subcarrier m of user k, and h_{km} is the corresponding channel gain which can be obtained by channel estimation [21]. Then, the channel condition of user k is given by the vector

$$\mathbf{h}_k = \mathrm{SNR}_k\left[|h_{k1}|^2, \ldots, |h_{kM}|^2\right]. \tag{4}$$

The channel condition of the whole system is given by $\mathbf{h} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$, and its state space is denoted by $\mathcal{H}$. We also let $\mathbf{b}_k = [\beta_{k1}, \ldots, \beta_{kM}]$, $\mathbf{b} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$, and $\mathcal{B}$ denote its state space.

In the interested OFDMA system, a timeslot is divided into multiple parts which will be allocated to the traffic of different type in each user. Let d_{kt} denote the generic traffic in $\mathcal{D}_{kt}$, which is the set of traffic for type $t \in \mathcal{T}$ in user $k \in \mathcal{K}$. Let $\alpha_{d_{kt}m}$ be the timeslot occupancy ratio of the subcarrier m for the traffic d_{kt}. Similar to the channel conditions of the OFDMA system, we let $\mathbf{a}_{d_{kt}} = [\alpha_{d_{kt}1}, \ldots, \alpha_{d_{kt}M}]$, $\mathbf{a} = [\mathbf{a}_{1_{11}}, \ldots, \mathbf{a}_{D_{KT}}]$, and $\mathcal{A}$ denote its state space. Thus, the transmission rate of traffic d_{kt} can be given by

$$r_{d_{kt}} = \sum_{m \in \mathcal{M}} \alpha_{d_{kt}m} C_{km}. \tag{5}$$

As stated in last subsection, there is no explicit QoS requirement for the first type of traffic $t_1 \in \mathcal{T}$. The QoS requirements of the second type of traffic $t_2 \in \mathcal{T}$ is Min R and Max R, which indicate that

$$R_{kt_2}^{\min} \le \mathbb{E}\left\{r_{d_{kt_2}}\right\} \le R_{kt_2}^{\max}, \tag{6}$$

where $r_{d_{kt_2}}$ can be calculated by (5). The QoS requirements of the third type of traffic $t_3 \in \mathcal{T}$ include Min R, Max R, and Max L, which indicate that

$$\begin{aligned} R_{kt_3}^{\min} &\le \mathbb{E}\left\{r_{d_{kt_3}}\right\} \le R_{kt_3}^{\max}, \\ l_{d_{kt_3}} &\le L_{kt_3}, \end{aligned} \tag{7}$$

where $l_{d_{kt_3}}$ is the latency of the traffic d_{kt_3}. In the timeslotted system, we have

$$l_{d_{kt_3}} = n \cdot \Delta + \varepsilon, \quad n \in \mathbb{N}, \tag{8}$$

where Δ is the length of timeslot and $0 \le \varepsilon < \Delta$. The QoS requirements of the fourth type of traffic $t_4 \in \mathcal{T}$ include Min R, Max R, Max L, and Tol J, which indicate that

$$\begin{aligned} R_{kt_4}^{\min} &= \mathbb{E}\left\{r_{d_{kt_4}}\right\} = R_{kt_4}^{\max}, \\ l_{d_{kt_4}} &\le L_{kt_4}, \\ j_{d_{kt_4}} &\le J_{kt_4}, \end{aligned} \tag{9}$$

where $l_{d_{kt_4}}$ has a similar relationship as (8), and $j_{d_{kt_4}}$ is the jitter of the traffic d_{kt_4}. According to the definition, $j_{d_{kt_4}}$ is given by

$$j_{d_{kt_4}} = \max_{\forall d'_{kt_4} \prec d_{kt_4}} \left| l_{d_{kt_4}} - l_{d'_{kt_4}} \right|, \tag{10}$$

where "$\prec$" denotes d'_{kt_4} was transmitted before d_{kt_4}.

3. Optimal Scheduling Policy

3.1. Cross-Layer Optimization Problem. The scheduling policy for this OFDMA system should transmit all the traffic as soon as possible, while guaranteeing the differential QoS requirements. As a cross-layer design problem, maximizing the spectrum efficiency is also an important consideration. Thus, we need to design a proper objective function to collect all the considerations. Similar to the methods in [9, 10, 13], the integrated residual workload is defined as follows.

Definition 1. Let $\mathcal{D}_{kt}$ be the set of traffic for type $t \in \mathcal{T}$ in user $k \in \mathcal{K}$ and $f(x)$ be a continuous strictly increasing nonnegative convex function for $x \ge 0$ and $f(0) = 0$. The integrated residual workload F at the end of the current timeslot is defined as

$$F = \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} \sum_{d_{kt} \in \mathcal{D}_{kt}} \kappa_{d_{kt}} \eta_{d_{kt}} f\left(d_{kt} - \Delta \cdot r_{d_{kt}}\right), \tag{11}$$

where Δ is the length of timeslot, $r_{d_{kt}}$ is the transmission rate allocated to traffic d_{kt}. $\kappa_{d_{kt}}$ is the function of the jitter $j_{d_{kt}}$, and $\eta_{d_{kt}}$ is the function of the latency $l_{d_{kt}}$. They are both the continuous strictly increasing nonnegative convex function, and they satisfy: (1) if $j_{d_{kt}} = 0, l_{d_{kt}} = 0$, then $\kappa_{d_{kt}} = 1, \eta_{d_{kt}} = 1$; (2) if $j_{d_{kt}} \to J_{kt}, l_{d_{kt}} \to L_{kt}$, then $\kappa_{d_{kt}} \to \infty, \eta_{d_{kt}} \to \infty$.

In this definition, $d_{kt} - \Delta \cdot r_{d_{kt}}$ is the residual workload of the traffic d_{kt} at the end of the current timeslot. Since the resource is allocated according to the transmission request, then we have $d_{kt} - \Delta \cdot r_{d_{kt}} \ge 0$. Here, $f(x)$ may have the form of x^2 according to its definition. It represents the punishment to the residual traffic in the queue. Clearly, $f(x)$ is increasing since there must be a greater punishment for more residual traffic. It can be seen that if $d_{kt} - \Delta \cdot r_{d_{kt}}$ is small, the small increase will not affect the stability of the scheduling system, that is, $f'(x)$ should be small at this time. However, if $d_{kt} - \Delta \cdot r_{d_{kt}}$ is large, a small increase may make the system unstable, that is, $f'(x)$ should be large. Thus, $f(x)$ must be a convex function when $x \ge 0$. $\kappa_{d_{kt}}$ and $\eta_{d_{kt}}$ represent the punishment to the jitter and the latency, respectively. According to their properties,

$$g(x) = \exp\left\{\frac{\psi x}{\xi - x}\right\}, \quad \psi > 0,\ 0 \le x < \xi \tag{12}$$

can satisfy the conditions in Definition 1, where ψ is the shape factor and ξ is the location parameter, which will be set to L or J. Thus, the integrated residual workload represents the residual workload of four types and their QoS requirements of delay and jitter. Thus, the cross-layer scheduling algorithm proposed in this paper is to minimize the integrated residual workload.

Before constructing the cross-layer optimization problem, we may do some preprocess on d_{kt} in order to simplify the problem. Note that the purpose of the maximum transmission rate is to restrict some greedy traffic to occupy too much bandwidth. Thus, if we do some operations on d_{kt} to make the transmission rate cannot be greater than $R_{kt}^{\max}$, then a group of constraints can be eliminated. Let $\widetilde{d}_{kt}$ be the transmission request after preprocess, then for every $t \in \mathcal{T}$ and $k \in \mathcal{K}$, we have

$$\widetilde{d}_{kt} = d_{kt} \mathbf{I}_{R_{kt}^{\max}}(d_{kt}) + \Delta \cdot R_{kt}^{\max}\left[1 - \mathbf{I}_{R_{kt}^{\max}}(d_{kt})\right], \tag{13}$$

where $\mathbf{I}_{R_{kt}^{\max}}(d_{kt})$ is the indicator function, which is defined as

$$\mathbf{I}_{R_{kt}^{\max}}(d_{kt}) = \begin{cases} 1, & d_{kt} \le \Delta \cdot R_{kt}^{\max}, \\ 0, & d_{kt} > \Delta \cdot R_{kt}^{\max}. \end{cases} \tag{14}$$

On the other hand, except for the type of traffic t_4, other three types are burst traffic. Thus, at the beginning of some timeslot, the traffic transmission request $\widetilde{d}_{kt}$ may be smaller than $\Delta \cdot R_{kt}^{\min}$. Then, we need to do some operations on $R_{kt}^{\min}$ in order to eliminate this contradiction. Let $\widetilde{R}_{kt}^{\min}$ be the minimum rate after preprocess, then for every $t \in \mathcal{T}$ and $k \in \mathcal{K}$, we have

$$\widetilde{R}_{kt}^{\min} = \frac{\widetilde{d}_{kt}}{\Delta}\mathbf{I}_{R_{kt}^{\min}}\left(\widetilde{d}_{kt}\right) + R_{kt}^{\min}\left[1 - \mathbf{I}_{R_{kt}^{\min}}\left(\widetilde{d}_{kt}\right)\right]. \tag{15}$$

Finally, collecting the scheduling objectives, QoS requirements, and physical constraints together, we have the following optimization problem:

$$\begin{aligned}
\min \quad & F = \sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}\sum_{\widetilde{d}_{kt}\in\mathcal{D}_{kt}} \kappa_{\widetilde{d}_{kt}}\eta_{\widetilde{d}_{kt}} f\left(\widetilde{d}_{kt} - \Delta \cdot r_{\widetilde{d}_{kt}}\right), \\
\text{s.t.} \quad & G_{\widetilde{d}_{kt_i}} = \widetilde{R}_{\widetilde{d}_{kt_i}}^{\min} - \overline{r}_{\widetilde{d}_{kt_i}}^{(n\Delta)} \le 0, \quad i = 2,3,4, \\
& G_{m+D} = \sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}\sum_{\widetilde{d}_{kt}\in\mathcal{D}_{kt}} \alpha_{\widetilde{d}_{kt}m} - 1 \le 0, \\
& G_{k+M+D} = \sum_{m\in\mathcal{M}} \beta_{km} - 1 \le 0, \\
& 0 \le \alpha_{\widetilde{d}_{kt}m} \le 1; \quad 0 \le \beta_{km} \le 1, \\
& \forall \widetilde{d}_{kt} \in \mathcal{D}_{kt}, \quad \forall t \in \mathcal{T}, \quad \forall k \in \mathcal{K}, \quad \forall m \in \mathcal{M},
\end{aligned} \tag{16}$$

where $D = \sum_{k\in\mathcal{K}}\sum_{i=2}^{4}|D_{kt_i}|$. In this formulation, F is the integrated residual workload after this time of traffic transmission. The constraints on $\alpha_{\widetilde{d}_{kt}m}$ means one subcarrier can be shared by all the traffic, while the constraint on β_{km} means, for each user, the sum of the power allocated to all subcarriers cannot exceed the total power constraint. If the traffic does not have a specific QoS requirement, the weighted function will be set to 1. The time average value of $r_{\widetilde{d}_{kt}}$ at epoch $n\Delta$, denoted by $\overline{r}_{\widetilde{d}_{kt}}^{(n\Delta)}$, is calculated as an exponentially weighted low-pass filter [22],

$$\overline{r}_{\widetilde{d}_{kt}}^{(n\Delta)} = \left(1 - \frac{1}{n}\right)\overline{r}_{\widetilde{d}_{kt}}^{((n-1)\Delta)} + \frac{1}{n}r_{\widetilde{d}_{kt}}. \tag{17}$$

3.2. Problem Decomposition. Equation (16) represents a complicated nonlinear optimization problem. In this section, we will propose a method to solve this problem with low complexity. Firstly, the following theorem shows the problem represented by (16) is convex.

Theorem 2. *The problem represented by* (16) *is a convex optimization problem, whose solution can be given by*

$$(\mathbf{a}^*, \mathbf{b}^*) = \arg\max_{\mathbf{a}\in\mathcal{A},\mathbf{b}\in\mathcal{B}}\left\{F + \sum_{i=1}^{K+M+D}\lambda_i G_i\right\}, \tag{18}$$

where λ_i is the Lagrangian multiplier, and $G_i < 0 \Rightarrow \lambda_i = 0$.

Proof. Consider the definition of convex optimization problem in [23]. First, the feasible region of the optimization variables $\alpha_{\widetilde{d}_{kt}m}$ and β_{km} constructs a convex polyhedron. Then, besides two groups of linear constraints, there are three groups of nonlinear constraints. Since a nonnegative weighted sum of convex functions is a convex function [23], then $\overline{r}_{\widetilde{d}_{kt}}^{(n\Delta)}$ is a concave function of $\alpha_{\widetilde{d}_{kt}m}$ and β_{km} according to (1), (3), and (5). Since $f(x)$ is an increasing convex function, $f(\widetilde{d}_{kt} - \Delta \cdot r_{\widetilde{d}_{kt}})$ is a convex function. Note that $\kappa_{\widetilde{d}_{kt}}$ and $\eta_{\widetilde{d}_{kt}}$ are constants, for the delay and the jitter are known, then F is a convex function. Since this is a convex optimization problem, the solutions expressed in (18) can be derived from Karush-Kuhn-Tucker (KKT) condition directly. □

Although the optimization problem represented by (16) is convex, the numerical algorithm for this problem still has a high computation complexity [23]. In the following, we will decompose this problem. The resulting problem enjoys a low complexity at a cost of trivial performance loss.

It should be noted that the layered optimization does not make big difference in terms of ergodic capacity [15]. Thus, we can decompose this problem into two steps: first, allocate subcarrier and timeslot to each type of traffic for every user; second, allocate power by using water-filling algorithm in each user. Since there are many works on the iterative implementation for water-filling [21], we only discuss the first step in detail. By using the equal power allocation and the quadratic objective function, the problem represented by (16) can be reduced to (19).

$$\begin{aligned}
\min \quad & F = \sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}\sum_{\widetilde{d}_{kt}\in\mathcal{D}_{kt}} \kappa_{\widetilde{d}_{kt}}\eta_{\widetilde{d}_{kt}}\left(\widetilde{d}_{kt} - \Delta \cdot r_{\widetilde{d}_{kt}}\right)^2, \\
\text{s.t.} \quad & G_{\widetilde{d}_{kt_i}} = \widetilde{R}_{\widetilde{d}_{kt_i}}^{\min} - \overline{r}_{\widetilde{d}_{kt_i}}^{(n\Delta)} \le 0, \quad i = 2,3,4, \\
& G_{m+D} = \sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}\sum_{\widetilde{d}_{kt}\in\mathcal{D}_{kt}} \alpha_{\widetilde{d}_{kt}m} - 1 \le 0, \\
& 0 \le \alpha_{\widetilde{d}_{kt}m} \le 1, \quad \forall \widetilde{d}_{kt} \in \mathcal{D}_{kt}, \\
& \forall t \in \mathcal{T}, \quad \forall k \in \mathcal{K}, \quad \forall m \in \mathcal{M}.
\end{aligned} \tag{19}$$

The resulting optimization problem in (19), referred to as the time-frequency allocation problem, is fortunately a continuous quadratic knapsack problem with a generalized upper bound and an angular structure in the constraints. The knapsack problem (integer or continuous) has been studied for decades, which has often been used to solve resource allocation problem in operational research, economics, military, and communications [16, 17]. According to the results in [16], we first form a Lagrangian relaxation with respect to the constraints $G_{m+D}, m = 1, \ldots, M$. The resulting Lagrangian subproblems then construct D singly constrained convex problems, that is,

$$\begin{aligned}
\min \quad & F_{d_{kt}} = \kappa_{\widetilde{d}_{kt}}\eta_{\widetilde{d}_{kt}}\left(\widetilde{d}_{kt} - \Delta \cdot r_{\widetilde{d}_{kt}}\right)^2 - \lambda\left(\sum_{\widetilde{d}_{kt}\in\mathcal{D}_{kt}} \alpha_{\widetilde{d}_{kt}m} - 1\right), \\
\text{s.t.} \quad & \widetilde{R}_{\widetilde{d}_{kt}}^{\min} - \overline{r}_{\widetilde{d}_{kt}}^{(n\Delta)} \le 0, \\
& 0 \le \alpha_{\widetilde{d}_{kt}m} \le 1.
\end{aligned} \tag{20}$$

(1) Receive the transmission request d_{kt}, $k \in \mathcal{K}$, $t \in$ and the QoS parameters.
(2) **for** $k \in \mathcal{K}$ and $t \in \mathcal{T}$ **do**
(3) **if** $d_{kt} > \Delta \cdot R_{kt}^{\max}$ **then**
(4) $d_{kt} \leftarrow \Delta \cdot R_{kt}^{\max}$.
(5) **else if** $d_{kt} < \Delta \cdot R_{kt}^{\min}$ **then**
(6) $R_{kt}^{\min} \leftarrow d_{kt}/\Delta$.
(7) **end if**
(8) **end for**
(9) Solve the optimization problem represented by (19).
(10) Transmit $\mathbf{a}^*$ to every user.

Algorithm 1: IRWM algorithm.

By using the vector $\boldsymbol{\alpha}_{d_{kt}}$, this problem can be converted into the following form

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\alpha}_{d_{kt}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\alpha}_{d_{kt}} + \mathbf{q}^{\mathrm{T}}\boldsymbol{\alpha}_{d_{kt}} + \lambda\mathbf{r}^{\mathrm{T}}\boldsymbol{\alpha}_{d_{kt}}, \\ \text{s.t.} \quad & \mathbf{e}^{\mathrm{T}}\boldsymbol{\alpha}_{d_{kt}} \geq 1, \quad 0 \leq \alpha_{\widetilde{d}_{kt}m} \leq 1. \end{aligned} \tag{21}$$

According to the algorithm proposed in [18, 19], this subproblem can be numerically solved efficiently.

3.3. Asymptotic Optimal Scheduling Policy. The feasible region of the problem represented by (19) might be an empty set, which means that the system may be unstable for some traffic transmission request and QoS requirements. The scheduling algorithm under which the system is stable is referred to as the stable scheduling algorithm (SSA). In order to discuss the stability of the scheduling algorithm, we define the static service split (SSS) scheduling algorithm which is similar to [9].

Definition 3. For every channel state $\mathbf{h} \in \mathcal{H}$, there is a fixed continuous probability measure $p(\mathbf{a}, \mathbf{b} \mid \mathbf{h})$, where $\mathbf{a} \in \mathcal{A}$ is the timeslot allocation vector and $\mathbf{b} \in \mathcal{B}$ is the power allocation vector. The SSS scheduling algorithm parameterized by the set of measures $\mathcal{P} \triangleq \{p(\mathbf{a}, \mathbf{b} \mid \mathbf{h}) : \mathbf{h} \in \mathcal{H}\}$. The average (or the long-term) service rate of traffic type $t \in \mathcal{T}$ in user $k \in \mathcal{K}$ is

$$\mathbb{E}\left\{r_{\widetilde{d}_{kt}}\right\} = \int_{\mathbf{h}} p(\mathbf{h}) \left(\int_{\mathbf{a}} \int_{\mathbf{b}} p(\mathbf{a}, \mathbf{b} \mid \mathbf{h}) r_{\widetilde{d}_{kt}} \, d\mathbf{a} \, d\mathbf{b} \right) d\mathbf{h}. \tag{22}$$

Then, $\mathcal{P}$ is called the SSS algorithm.

Similar to [9], the simple observation shows that if $F < \infty$ and the constrains $G_{\widetilde{d}_{kt_i}}$ hold, then the SSS algorithm, allocating to each traffic the average rate, will make the system stable. This fact gives the condition on which the system is stable.

Lemma 4. *Let $\widetilde{R}_{kt_i}^{\min}, i = 2, 3, 4$ be the minimum reserved rate, and $L_{kt_i}, i = 3, 4$, J_{kt_4} are the maximum latency and tolerant jitter, respectively. The sufficient condition for the existence of a SSA is for at least one SSS algorithm, the integrated residual workload F exists, and the following equations hold for every $\widetilde{d}_{kt} \in \mathcal{D}_{kt}, k \in \mathcal{K}, t \in \mathcal{T}$,*

$$\widetilde{R}_{kt_i}^{\min} \leq \mathbb{E}\left\{r_{\widetilde{d}_{kt_i}}\right\}, \quad i = 2, 3, 4. \tag{23}$$

From this lemma, one can define the scheduling algorithm stability region $\mathcal{R}$ as the QoS requirements set which satisfies Lemma 4. Then, the asymptotic properties of the optimization problem represented by (19) can be summarized as the following theorem.

Theorem 5. *If QoS parameters are in the scheduling algorithm stability region $\mathcal{R}$, then the solution of the optimization problem represented by* (19) *satisfies the QoS requirements of* (6), (7), *and* (9) *when $n \to \infty$, and minimizes the integrated residual workload F.*

Proof. If the QoS requirements are in the region $\mathcal{R}$, according to Lemma 4, the SSA must exist. So, the feasible domain of the optimization problem represented by (19) is not null. According to Theorem 2, the optimal solution of the problem represented by (19) exists. Because the arrival rate of traffic $t_4 \in \mathcal{T}$ is $\widetilde{R}_{kt_4}^{\min}$, which is also the requesting rate, then $\overline{r}_{\widetilde{d}_{kt_4}}^{(n\Delta)}$ is equal to $\widetilde{R}_{kt_4}^{\min}$ as long as the optimal solution exists. According to the law of large numbers, the average rates in time are equal to their mathematical expectations, then (6), (7), and (9) hold. □

The scheduling algorithm executes as in Algorithm 1: users offer traffic transmission requests and QoS parameters at the beginning of each timeslot, meanwhile the BS estimates the uplink wireless channel condition, then the BS solves the problem represented by (19) and sends the resource allocation results to all users. After receiving $\mathbf{a}^*$, each user executes the water-filling algorithm independently to obtain $\mathbf{b}^*$. As this algorithm always tries to minimize the integrated residual workload, it will be referred to as the *integrated residual workload minimization* (IRWM) algorithm.

3.4. Heuristic Call Admission Control. For an OFDMA system in the heavily loaded scenario, the stability of the queues cannot always be assured. In this case, the optimization problem represented by (19) will have a null feasible region.

(1) Determine $\widetilde{R}_{kt}^{\min}$, $R_{kt}^{\max}$, L_{kt} and J_{kt} for a specific $k \in \mathcal{K}$ and $t \in \mathcal{T}$.
(2) Add $\widetilde{R}_{kt}^{\min}$, L_{kt} and J_{kt} to (19).
(3) $l_{\widetilde{d}_{kt}} \leftarrow 0$, $j_{\widetilde{d}_{kt}} \leftarrow 0$.
(4) $\widetilde{d}_{kt} \leftarrow \Delta \cdot R_{kt}^{\min}, \forall k \in \mathcal{K}, \forall t \in \mathcal{T}$.
(5) **if a*** exists **then**
(6) Admit.
(7) **else**
(8) Reject.
(9) **end if**

Algorithm 2: Heuristic CAC algorithm.

Table 1: Parameters of the traffic sources for two users.

Traffic source	Type t_1	Type t_2	Type t_3	Type t_4
ON state length	$EXP(10)$	∞	∞	∞
OFF state length	$EXP(10)$	0	0	0
Interarrival time	$EXP(0.25)$	$EXP(0.25)$	$EXP(0.25)$	1
Packet size	$EXP(100)$	$EXP(100)$	$EXP(100)$	200

Table 2: QoS parameters of each traffic type for two users.

QoS parameters	Type t_1	Type t_2	Type t_3	Type t_4
Min R	–	100	100	200
Max R	–	300	300	200
Max L	–	–	1.5	1
Tol J	–	–	–	0.5

To overcome this problem, we need to design a call admission control (CAC) mechanism. The algorithm based on this idea is listed as Algorithm 2. Join this heuristic CAC algorithm and the IRWM algorithm will form a cross-layer resource allocation and scheduling framework for OFDMA wireless networks supporting multiple types of traffic.

4. Simulation Results

The uplink scenario of one BS and 8 users is addressed in this section. The wireless channel between each user and the base station undergoes 16-path frequency selective fading. The OFDMA system considered has 256 subcarriers, and the bandwidth for each subcarrier is 50 Hz. The channel gains for different subcarriers are independent and identical distribution and the variance is 1. The average SNR for the first four users are 20 dB and 10 dB for the second user. The target BER of AMC mechanism is 10^{-4}. If we allocate transmission power equally, then the channel capacity is about 687 bit/s for the first four users and about 546 bit/s for the second four users. We consider the time duration of 1,000 timeslots.

The ON-OFF model is used to generate the traffic for each user. The traffic parameters are listed in Table 1, where $EXP(\lambda)$ is the exponential distribution with the average λ. The total average arrival rate is 600 bit/s, which is bigger than the channel capacity of the second group of users with equal power allocation. The QoS requirements are shown in Table 2. In these tables, the time unit is the length of timeslot Δ, the traffic unit is bit and the transmission rate unit is bit/timeslot. In the objective function, we let $f(x)$ be x^2. The weighted functions for the latency and the jitter have the form as (12), whose shape parameters are the Max L and Tol J, respectively.

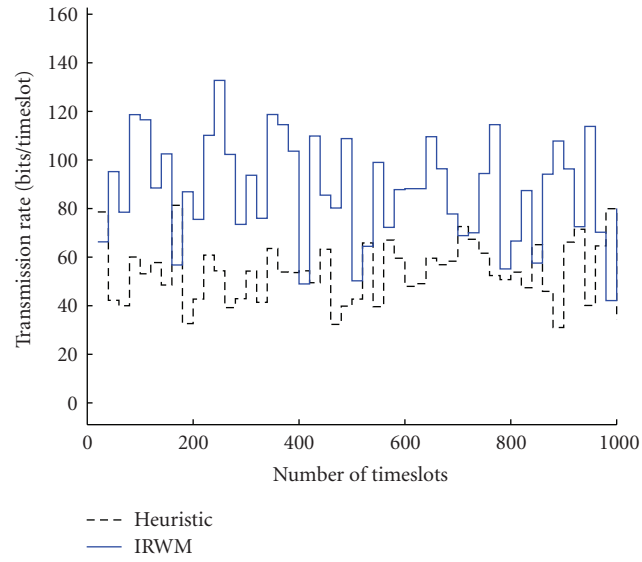

Figure 2: Transmission rate of traffic type t_1.

The simulation results for the second user are shown in Figures 2–7. From Figures 2–5, we can see that the average transmission rate is greater than the minimum rate or equal to the constant rate. So, the IRWM algorithm can guarantee the minimum reserved rate requirements. Figure 6 shows the latency of traffic type t_3. The largest traffic latency is about 1.45, it does not exceed the maximum latency requirement 1.5. The latency of traffic type t_4 is shown in Figure 7, which does not exceed the corresponding maximum value in Table 2 too. So, the IRWM algorithm can guarantee the maximum latency and the tolerant jitter requirements.

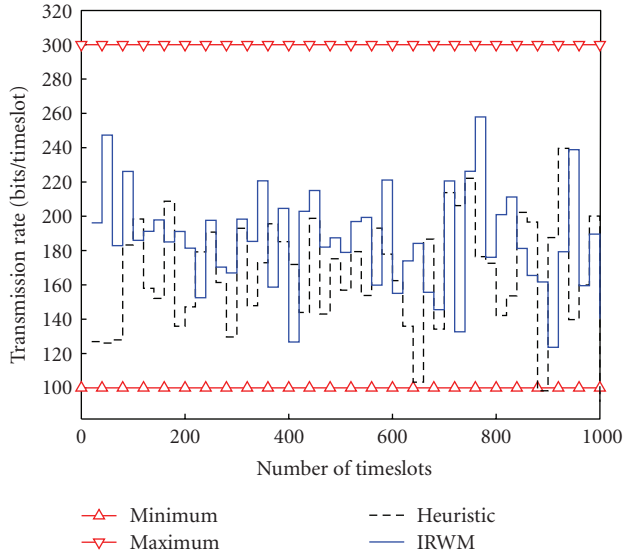

Figure 3: Transmission rate of traffic type t_2.

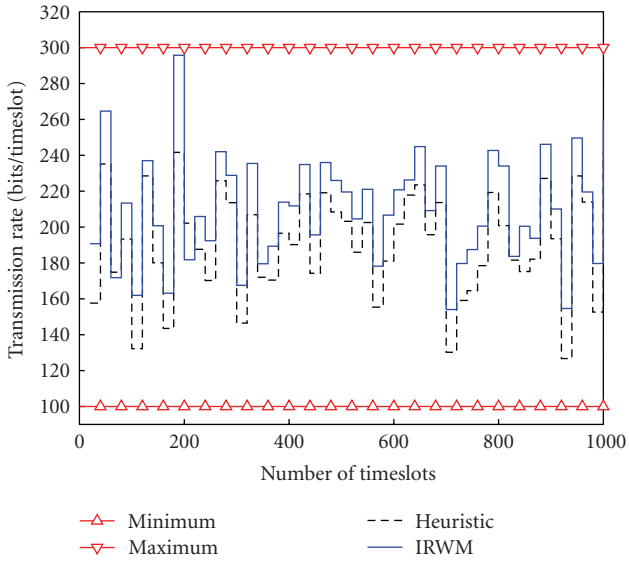

Figure 4: Transmission rate of traffic type t_3.

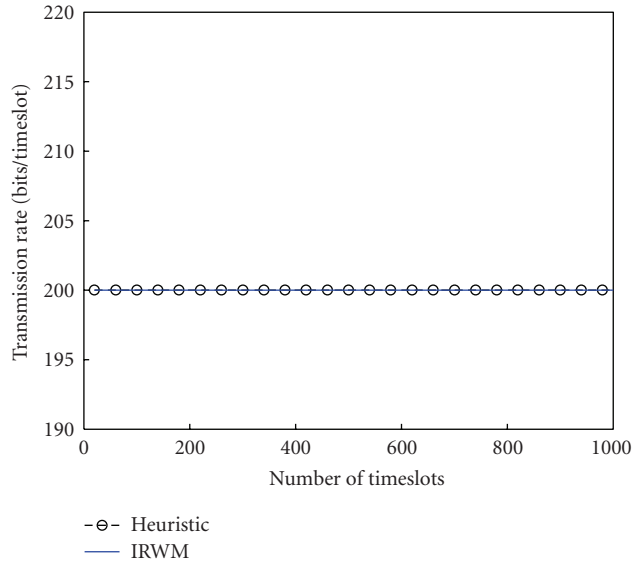

Figure 5: Transmission rate of traffic type t_4.

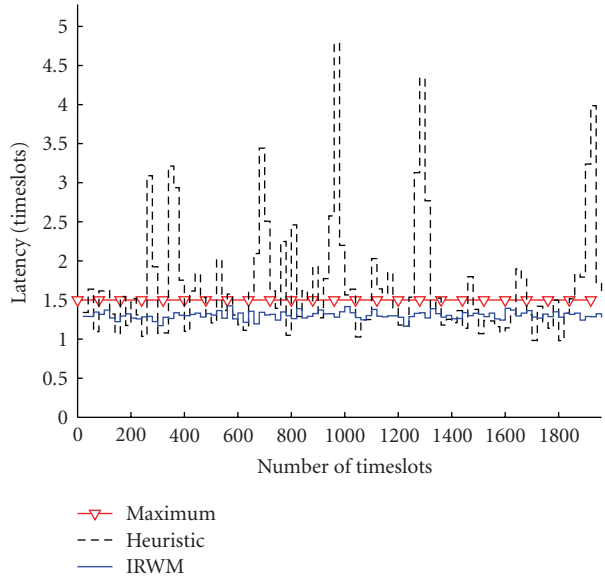

Figure 6: Latency of traffic type t_3.

For performance comparison, the heuristic scheme has also been simulated. In this scheme, the interleaved subcarrier allocation is used. The subcarriers are allocated to the traffic of type t_4 first. Then, according to the traffic requirements and QoS parameters, the subcarriers are allocated to the traffic of types t_3 and t_2, respectively. At last, the residual subcarriers are allocated to the traffic of type t_1. In this scheme, the maximum sustainable rates of traffic types t_3 and t_2 are two critical parameters, which balance the transmission among traffic types t_3, t_2, and traffic type t_1. If the maximum sustainable rate is too large, the traffic of type t_1 can nearly not get transmission opportunities, while if it is too small, the latency requirement of traffic types t_3 will be violated. In IRWM algorithm; however, there is no need to set the maximum sustainable rate manually, because the integrated residual workload can balance all the types of traffic automatically. The simulation results show that the proposed IRWM algorithm has a better performance. It has a greater transmission rate for traffic types of t_1, t_2, and t_3. It also yields a smaller latency for the traffic type of t_1. Therefore, the simulation results show that the differential QoS requirements of four types of traffic are guaranteed effectively by the proposed IRWM algorithm.

5. Conclusion

The problem of uplink traffic scheduling with differential QoS requirements in OFDMA systems was addressed in

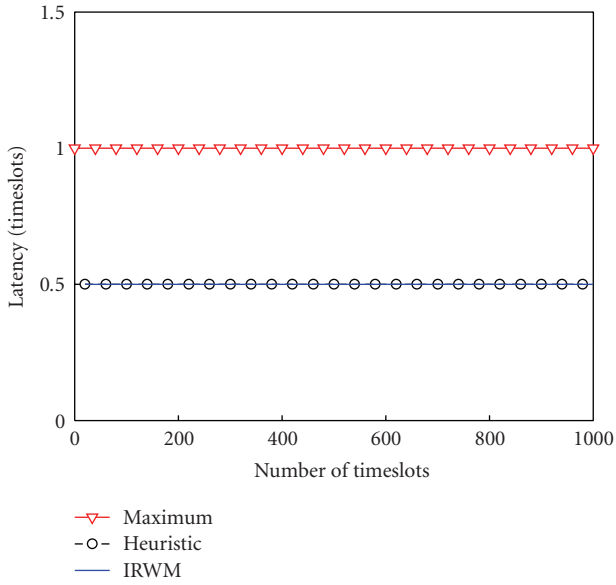

Figure 7: Latency of traffic type t_4.

this paper. A cross-layer optimization methodology, which jointly considers the traffic arrival process and the wireless channel conditions, was adopted to achieve better QoS for the users accessing to a common base station. In particular, we introduce the integrated residual workload to formulate the traffic scheduling problem into a convex optimization problem. By decomposing this problem into two steps, that is, a continuous quadratic knapsack problem in BS and a water-filling power allocation algorithm in each user, we presented a low-complexity algorithm referred to as the IRWM. Besides, a heuristic CAC scheme was proposed to avoid the sharply decreasing of QoS, when the system is in congestion. Both the theoretical analysis and the simulation results showed that the differential QoS requirements of the application layer are guaranteed effectively by the proposed algorithm in the MAC layer.

Acknowledgment

This work is supported by NSFC key project under Grant no. 60832008, and RGC/NSFC project under Grant no. N_HKUST622/06.

References

[1] S. H. Ali, K.-D. Lee, and V. C. M. Leung, "Dynamic resource allocation in OFDMA wireless metropolitan area networks," *IEEE Wireless Communications*, vol. 14, no. 1, pp. 6–13, 2007.

[2] B. Bai, W. Chen, Z. Cao, and K. B. Letaief, "Max-matching diversity in OFDMA systems," *IEEE Transactions on Communications*, vol. 58, no. 4, pp. 1161–1171, 2010.

[3] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[4] S. Sesia, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution—From Theory to Practice*, John Wiley & Sons, New York, NY, USA, 2009.

[5] *IEEE Std. 802.16e^{TM}, IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1*, IEEE Press, New York, NY, USA, 2005.

[6] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX—Understanding Broadband Wireless Networking*, Prentice Hall, New York, NY, USA, 2007.

[7] I. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 963–969, 1995.

[8] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, 2004.

[9] A. L. Stolyar, "Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic," *Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, 2004.

[10] A. Mandelbaum and A. L. Stolyar, "Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule," *Operations Research*, vol. 52, no. 6, pp. 836–855, 2004.

[11] D. Niyato and E. Hossain, "A queuing-theoretic and optimization-based model for radio resource management in IEEE 802.16 broadband wireless networks," *IEEE Transactions on Computers*, vol. 55, no. 11, pp. 1473–1488, 2006.

[12] G. Song, Y. Li, L. J. Cimini Jr., and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '04)*, vol. 3, pp. 1939–1944, Atlanta, Ga, USA, March 2004.

[13] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling: generalizations to include multiple constraints, multiple interfaces, and short term fairness," *Wireless Networks*, vol. 11, no. 5, pp. 557–569, 2005.

[14] A. Fu and J. N. Tsitsiklis, "Optimal transmission scheduling over a fading channel with energy and deadline constraints," *IEEE Transactions on Wireless Communications*, vol. 5, no. 2, pp. 630–641, 2006.

[15] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, 1997.

[16] K. M. Bretthauer and B. Shetty, "The nonlinear knapsack problem—algorithms and applications," *European Journal of Operational Research*, vol. 138, no. 3, pp. 459–472, 2002.

[17] M. Patriksson, "A survey on the continuous nonlinear resource allocation problem," *European Journal of Operational Research*, vol. 185, no. 1, pp. 1–46, 2008.

[18] J.-S. Pang, "A new and efficient algorithm for a class of portfolio selection problems," *Operations Research*, vol. 28, no. 3, pp. 754–767,, 1980.

[19] K. M. Bretthauer and B. Shetty, "Quadratic resource allocation with generalized upper bounds," *Operations Research Letters*, vol. 20, no. 2, pp. 51–57, 1997.

[20] A. J. Goldsmith and S. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218–1230, 1997.

[21] D. Tse and D. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, New York, NY, USA, 2005.

[22] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.