

Resource Allocation in OFDMA Wireless Communications Systems Supporting Multimedia Services

Kae Won Choi, Wha Sook Jeon, *Senior Member, IEEE*, and Dong Geun Jeong, *Senior Member, IEEE*

Abstract—We design a resource allocation algorithm for down-link of orthogonal frequency division multiple access (OFDMA) systems supporting real-time (RT) and best-effort (BE) services simultaneously over a time-varying wireless channel. The proposed algorithm aims at maximizing system throughput while satisfying quality of service (QoS) requirements of the RT and BE services. We take two kinds of QoS requirements into account. One is the required average transmission rate for both RT and BE services. The other is the tolerable average absolute deviation of transmission rate (AADTR) just for the RT services, which is used to control the fluctuation in transmission rates and to limit the RT packet delay to a moderate level. We formulate the optimization problem representing the resource allocation under consideration and solve it by using the dual optimization technique and the projection stochastic subgradient method. Simulation results show that the proposed algorithm well meets the QoS requirements with the high throughput and outperforms the modified largest weighted delay first (M-LWDF) algorithm that supports similar QoS requirements.

Index Terms—Multimedia communications, orthogonal frequency division multiple access (OFDMA), quality of service (QoS), radio resource allocation, wireless network.

I. INTRODUCTION

SINCE orthogonal frequency division multiple access (OFDMA) systems can offer a high data rate for guaranteeing various quality of service (QoS) requirements to a large number of users, OFDMA is regarded as one of the most promising candidates for the multiple access technique of current and future wireless multimedia communications systems. In the OFDMA systems, radio resource is represented in both frequency and time domains and can be very flexibly allocated to the users according to their need. Thus, an accurate

resource allocation algorithm is essential to assure the inherent capabilities of the OFDMA system.

There have been plenty of studies in this area (e.g., [1]–[6]). However, most of them do not consider various traffic classes with different QoS requirements and time-varying channel conditions simultaneously, which generally allow the opportunistic resource allocation. Instead, they focus on maximizing the system throughput under the constraints on power and transmission rates [1], [2], or on minimizing the transmission power under the constraints on transmission rates [3]–[5]. Even though [6] proposes the opportunistic scheduling algorithm for the OFDMA system with time-varying channel, it supports only the QoS requirements for best-effort traffic and does not consider the coexistence with other traffic classes.

In [7], we have suggested a packet scheduling and resource allocation algorithm for real-time and non-real-time traffic. Although this algorithm deals with various traffic classes, there is no guarantee for the optimality since it has been designed by a heuristic approach. In this paper, we propose a resource allocation algorithm based on the dual optimization technique, which maximizes the OFDMA system throughput while satisfying the QoS requirements of both real-time (RT) and best-effort (BE) traffic over time-varying channel.

In the wireless systems with time-varying channels, the resource allocation algorithm can exploit channel variation to enhance the system performance. This concept of the opportunistic resource allocation has been widely applied in the packet schedulers for the third generation mobile communications systems [8]–[11]. Although this strategy improves the system throughput, it can cause starvation of the user who suffers from a bad channel for a long time, which results in the excessive packet delay. For the RT users, the excessive delay can lead to severe performance degradation. In the algorithm proposed in this paper, this difficulty is overcome by the restricted exploitation of channel variation, where the restriction is given by the QoS requirements of RT and BE traffic.

We consider the video streaming service as the representative RT service since it generates massive traffic in comparison with other multimedia services. Two kinds of QoS requirements for the RT (video) service are defined in this paper. The first is the “required average transmission rate,” which is usually set to the average source data rate of the video service. The second QoS requirement is on the variation in transmission rate. With the opportunistic resource allocation, the severe fluctuation of available transmission rate is likely to occur frequently, which can cause the excessive RT packet delay. If the fluctuation can be limited by assigning more resources to the users suffering from starvation even though their channel conditions are bad,

Manuscript received May 25, 2006; revised May 04, 2007 and December 06, 2007; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor N. Shroff. First published September 09, 2008; current version published June 17, 2009. This research was supported by the Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program supervised by the Institute for Information Technology Advancement (ITA) (ITA-2008-C1090-0803-0002).

K. W. Choi was with the School of Electrical Engineering and Computer Science, Seoul National University, Seoul 151-742, Korea. He is now with Telecommunication Business, Samsung Electronics, Suwon-city, Gyeonggi-do 443-742, Korea (e-mail: kaewon.choi@gmail.com).

W. S. Jeon is with the School of Electrical Engineering and Computer Science, Seoul National University, Seoul 151-742, Korea (e-mail: wsjeon@snu.ac.kr).

D. G. Jeong is with the School of Electronics and Information Engineering, Hankuk University of Foreign Studies, Yongin-si, Kyonggi-do 449-791, Korea (e-mail: dgjeong@hufs.ac.kr).

Digital Object Identifier 10.1109/TNET.2008.2001470

the probability of the excessive packet delay can be lessened. In this paper, we take the average absolute deviation of transmission rate (AADTR) as the measure for fluctuation of transmission rate and design the resource allocation algorithm to restrict AADTR to the predefined “tolerable AADTR.”

For the BE services, we consider the required average transmission rate as the QoS requirement, to prevent the long starvation of some users and the excessive delay of their packets. This policy is particularly helpful for the Internet services using the transmission control protocol (TCP), because the excessive delay for a user can cause the slow-start in the TCP congestion control mechanism and in turn it leads to the degradation in system performance [11], [12].

To design the resource allocation algorithm, we first formulate the optimization problem which maximizes the total average transmission rate of BE traffic under the constraints on the required average transmission rates of the RT and BE services and AADTR of the RT service. Then, we solve this problem by using the dual optimization technique [13] and the projection stochastic subgradient method [14].

We have utilized the dual optimization technique to design the packet transmission scheduler in [12], although the design concept and the structure of the scheduler are substantially different from those of the OFDMA resource allocator in this paper. It is also noted that the dual optimization technique has been used in [10] and [15]. The algorithms in [10] and [15] can support only the QoS of non-real-time traffic (e.g., fairness, minimum expected transmission rate) for the time-division multiple access (TDMA) and the code-division multiple access (CDMA) systems, respectively. On the contrary, the proposed algorithm is designed for the OFDMA systems and is able to support the RT and BE traffic simultaneously.

The rest of the paper is organized as follows. In Section II, we describe the system model. In Section III, we formulate the resource allocation problem and design the proposed resource allocation algorithm. Section IV presents the simulation results. Finally, the paper is concluded with Section V.

II. SYSTEM MODEL

We consider the downlink of an OFDMA system that consists of one base station (BS)¹ and N users. The users are indexed by $n (= 1, \dots, N)$. Each user is assumed to have only one class of traffic which is either RT or BE. A user having RT (BE) traffic is called an “RT user” (“BE user”). The numbers of RT and BE users are respectively denoted by N_{RT} and N_{BE} . Then, we have $N_{\text{RT}} + N_{\text{BE}} = N$. The RT users are indexed by $1, \dots, N_{\text{RT}}$, and the BE users are indexed by $N_{\text{RT}} + 1, \dots, N$.

Time is divided into frames indexed by t . Each frame contains B OFDM symbols, and the duration of an OFDM symbol is denoted by T_s . The system uses M subcarriers. Subcarriers are indexed by $m (= 1, \dots, M)$. A subcarrier can be allocated to only a user at a time.

We assume that the duration of a frame (i.e., $B \cdot T_s$) is shorter than the channel coherence time, and therefore channel gain remains constant in a frame. We define the “channel state” to represent the combination of the channel gains of all subcarriers

for all users. The channel state is indexed by $k \in \mathcal{K}$, where \mathcal{K} is the set of all possible channel states. We define $h_{k,n,m}$ be the complex channel gain of the subcarrier m for the user n when the channel state is k . The channel state changes from frame to frame. Let $\eta(t)$ denote the channel state at the frame t . If $\eta(t) = k$, the channel gain of the subcarrier m for the user n at the frame t is equal to $h_{k,n,m}$.

It is assumed that the total available power of the BS, denoted by P , is evenly distributed to all subcarriers for simplicity. Then, the energy assigned to a subcarrier in a symbol duration is PT_s/M . Let $\rho_{k,n,m}$ be defined as the achievable data rate at which the user n can receive data by the subcarrier m when the channel state is k . If we assume that the achievable data rate is equal to the Shannon capacity, we have

$$\rho_{k,n,m} = \frac{1}{T_s} \cdot \log_2 \left(1 + \frac{PT_s |h_{k,n,m}|^2}{MN_o} \right) \quad (1)$$

where N_o is the variance of a circular symmetric complex Gaussian noise.² It is assumed that the BS can calculate the achievable data rates of all subcarriers for all users, based on the signal-to-noise ratios (SNRs) reported from users.

Since the channel state changes over time, the achievable data rate also does. Let $\rho_{n,m}(t)$ denote the achievable data rate at which the user n can receive data by the subcarrier m at the frame t . If $\eta(t) = k$, we have $\rho_{n,m}(t) = \rho_{k,n,m}$ for all n and m . It is assumed that $\{\eta(t) | t = 1, \dots\}$ is a stationary random process, and the probability that $\eta(t) = k$ is p_k . In other words, $\{\rho_{n,m}(t) | t = 1, \dots\}$ is a stationary random process and the probability that $\rho_{n,m}(t) = \rho_{k,n,m}$ for all n and m is p_k .

The proposed resource allocation algorithm decides the transmission rate of each user every frame, based only on the current channel state. Let $r_{k,n}$ be a possible transmission rate of user n when the channel state is k . In this paper, we use a bold face to represent a vector (e.g., \mathbf{x}), and a bold face with a bar to represent a matrix (e.g., $\bar{\mathbf{x}}$). Let $\mathbf{r}_k := (r_{k,1}, \dots, r_{k,N})^T$ denote a possible transmission rate vector. And we define $\bar{\mathbf{r}} := (r_{k,n})_{\substack{k \in \mathcal{K} \\ n=1, \dots, N}}$ as a possible transmission rate matrix.

The transmission rate of user n depends on which subcarriers are allocated to the user n . Let $\psi_{n,m}$ be the subcarrier allocation indicator that is 1 only when the subcarrier m is allocated to the user n . Otherwise, $\psi_{n,m}$ is 0. Even though $\psi_{n,m}$ can be 0 or 1, we assume that $\psi_{n,m}$ can be any value satisfying $0 \leq \psi_{n,m} \leq 1$ for mathematical tractability. Since a subcarrier can be allocated to only a user at a time, the condition $\sum_{n=1}^N \psi_{n,m} \leq 1$ should be satisfied for all m . Let \mathcal{C}_k be the set of all possible transmission rate vectors when the channel state is k . Then \mathcal{C}_k is defined as

$$\mathcal{C}_k := \left\{ (r_{k,1}, \dots, r_{k,N})^T \mid r_{k,n} \leq \sum_{m=1}^M \rho_{k,n,m} \psi_{n,m}, \right. \\ \left. \sum_{n=1}^N \psi_{n,m} \leq 1, 0 \leq \psi_{n,m} \leq 1, \right. \\ \left. \forall n = 1, \dots, N, \forall m = 1, \dots, M \right\}. \quad (2)$$

¹In this paper, the term “BS” stands for the central controller in various wireless networks, for example, the BS in cellular mobile networks and the access point in wireless local area networks.

²When applying the proposed algorithm to the practical systems, we can redefine the achievable data rate to be more appropriate to the practical modulation and coding techniques.

From (2), we can deduce that \mathcal{C}_k is a convex, closed, and bounded set. We also define the set of all possible transmission matrixes, $\bar{\mathcal{C}}$, as

$$\bar{\mathcal{C}} := \{\bar{\mathbf{r}} | \mathbf{r}_k \in \mathcal{C}_k, \forall k \in \mathcal{K}\}. \quad (3)$$

The proposed algorithm selects a transmission rate vector every frame according to the current channel state. At frame t , if the channel state is k (i.e., $\eta(t) = k$), the algorithm select \mathbf{r}_k as the transmission rate vector for the frame, where \mathbf{r}_k should be within the set \mathcal{C}_k . The task of the resource allocation algorithm is to select the transmission rate vector \mathbf{r}_k out of the set \mathcal{C}_k for all channel states $k \in \mathcal{K}$ to maximize the system performance.

III. RESOURCE ALLOCATION ALGORITHM

A. QoS Requirements

The proposed resource allocation algorithm aims at maximizing the sum of the average transmission rates of BE users (thus, the total system throughput), while satisfying the QoS requirements that are the required average transmission rate and the tolerable AADTR for RT users and only the required average transmission rate for BE users.

We define R_n as the required average transmission rate for the user n , and $\mathbf{R} := (R_1, \dots, R_N)^T$. The traffic generation rate at an RT source (i.e., the source rate) is generally modeled as a variable bit rate. The required average transmission rate for the RT user should be configured as the long-term average of the traffic generation rate or, conservatively, slightly more than that. For the BE users, it is desirable that the required average transmission rate is set to a small value that can prevent starvation, that is, the minimum transmission rate. The constraints on the average transmission rates are expressed as follows:

$$\sum_{k \in \mathcal{K}} p_k r_{k,n} \geq R_n, \quad \text{for } n = 1, \dots, N. \quad (4)$$

These constraints on transmission rates are not enough to guarantee the QoS of RT users. As mentioned before, the large-scale fluctuation in transmission rate can incur the excessive packet transmission delay. We consider the tolerable AADTR as another QoS requirement for RT services to control the fluctuation and to limit the delay to a moderate level. Let D_n be the tolerable AADTR of user n , and $\mathbf{D} := (D_1, \dots, D_{N_{\text{RT}}})^T$. Then the constraints on AADTR are as follows:

$$\sum_{k \in \mathcal{K}} p_k |r_{k,n} - R_n| \leq D_n, \quad \text{for } n = 1, \dots, N_{\text{RT}}. \quad (5)$$

The throughput of BE users and the number of useless RT packets by the excessive delay increase together as the tolerable AADTR gets higher. To maximize the throughput, the tolerable AADTR should be set to the largest as long as the number of useless packets is allowable. In the practical system design, the appropriate value of the tolerable AADTR can be found by the field trials and/or the computer simulation.

B. Problem Formulation

The proposed resource allocation algorithm solves the following optimization problem to maximize the sum of

the throughputs of all BE users while satisfying the QoS requirements.

$$\begin{aligned} \max \quad & \sum_{n=N_{\text{RT}}+1}^N \sum_{k \in \mathcal{K}} p_k r_{k,n} \\ \text{s.t.} \quad & \sum_{k \in \mathcal{K}} p_k r_{k,n} \geq R_n, \quad \forall n = 1, \dots, N \\ & \sum_{k \in \mathcal{K}} p_k |r_{k,n} - R_n| \leq D_n, \quad \forall n = 1, \dots, N_{\text{RT}} \end{aligned} \quad (6)$$

where $\mathbf{r}_k \in \mathcal{C}_k$ for all $k \in \mathcal{K}$. Let $\mathbf{r}_k^* := (r_{k,1}^*, \dots, r_{k,N}^*)^T$ denote the solution to the problem (6). And we also define $\bar{\mathbf{r}}^* := (r_{k,n}^*)_{n=1, \dots, N, k \in \mathcal{K}}$. It is noted that the optimization problem (6) is convex.

We introduce the dual problem of (6) since it has the more favorable structure than the primal one. Let us define $\phi(\mathbf{r}_k) := (\alpha_1(\mathbf{r}_k), \dots, \alpha_N(\mathbf{r}_k), \beta_1(\mathbf{r}_k), \dots, \beta_{N_{\text{RT}}}(\mathbf{r}_k))^T$, where $\alpha_n(\mathbf{r}_k) := r_{k,n}$ for $n = 1, \dots, N$ and $\beta_n(\mathbf{r}_k) := -|r_{k,n} - R_n|$ for $n = 1, \dots, N_{\text{RT}}$. Then, the Lagrangian is

$$\begin{aligned} L(\bar{\mathbf{r}}, \mathbf{w}) &:= \sum_{n=N_{\text{RT}}+1}^N \sum_{k \in \mathcal{K}} p_k r_{k,n} \\ &+ \sum_{n=1}^N u_n \left(\sum_{k \in \mathcal{K}} p_k r_{k,n} - R_n \right) \\ &+ \sum_{n=1}^{N_{\text{RT}}} v_n \left(D_n - \sum_{k \in \mathcal{K}} p_k |r_{k,n} - R_n| \right) \\ &= \sum_{k \in \mathcal{K}} p_k (\mathbf{e} + \mathbf{w})^T \phi(\mathbf{r}_k) - \mathbf{u}^T \mathbf{R} + \mathbf{v}^T \mathbf{D} \end{aligned} \quad (7)$$

where u_n and v_n are the Lagrange multipliers, $\mathbf{u} := (u_1, \dots, u_N)^T$, $\mathbf{v} := (v_1, \dots, v_{N_{\text{RT}}})^T$, and $\mathbf{w} := (u_1, \dots, u_N, v_1, \dots, v_{N_{\text{RT}}})^T$. In (7), $\mathbf{e} := (e_1, \dots, e_{N+N_{\text{RT}}})^T$, where $e_n = 1$ for $n = N_{\text{RT}} + 1, \dots, N$ and $e_n = 0$ for the other n 's.

The Lagrange dual function is

$$\begin{aligned} g(\mathbf{w}) &:= \max_{\bar{\mathbf{x}} \in \bar{\mathcal{C}}} L(\bar{\mathbf{x}}, \mathbf{w}) \\ &= \sum_{k \in \mathcal{K}} p_k \max_{\mathbf{x} \in \mathcal{C}_k} (\mathbf{e} + \mathbf{w})^T \phi(\mathbf{x}) - \mathbf{u}^T \mathbf{R} + \mathbf{v}^T \mathbf{D}. \end{aligned} \quad (8)$$

The dual problem is

$$\begin{aligned} \min \quad & g(\mathbf{w}) \\ \text{s.t.} \quad & \mathbf{w} \succeq \mathbf{0} \end{aligned} \quad (9)$$

where $\mathbf{0}$ is the vector of which all components are 0, and the notation \succeq is a component wise inequality, that is, when $\mathbf{x} = (x_1, \dots, x_N)^T$ and $\mathbf{y} = (y_1, \dots, y_N)^T$, $\mathbf{x} \succeq \mathbf{y}$ if and only if $x_n \geq y_n$ for all n . Let \mathcal{W}^* denote the set of all solutions of (9), and \mathbf{w}^* denote one of the solutions, i.e., $\mathbf{w}^* \in \mathcal{W}^*$.

C. Transmission Rate Decision

Let us define $\mathcal{S}_k(\mathbf{w})$ as follows.

$$\mathcal{S}_k(\mathbf{w}) := \arg \max_{\mathbf{x} \in \mathcal{C}_k} (\mathbf{e} + \mathbf{w})^T \phi(\mathbf{x}). \quad (10)$$

For deriving the solutions of the primal problem and dual problem, it is needed to compute the transmission rate vector

$\mathbf{s}_k(\mathbf{w}) := (s_{k,1}(\mathbf{w}), \dots, s_{k,N}(\mathbf{w}))^T$ such that $\mathbf{s}_k(\mathbf{w}) \in \mathcal{S}_k(\mathbf{w})$ at each frame with channel state k .

We define $\bar{\mathbf{s}}(\mathbf{w}) = (\mathbf{s}_k(\mathbf{w}))_{k \in \mathcal{K}}$. Then, we have $\bar{\mathbf{s}}(\mathbf{w}) \in \arg \max_{\bar{\mathbf{x}} \in \bar{\mathcal{C}}} L(\bar{\mathbf{x}}, \mathbf{w})$. The use of this vector is twofold. First, it is used to get the solution of the dual problem. Second, it in itself is the solution of the primal problem when $\mathbf{w} \in \mathcal{W}^*$.

The vector $\mathbf{s}_k(\mathbf{w})$ can be found by solving the following convex optimization problem:

$$\begin{aligned} \max \quad & (\mathbf{e} + \mathbf{w})^T \boldsymbol{\phi}(\boldsymbol{\gamma}) = \sum_{n=1}^N U_n(\gamma_n, \mathbf{w}) \\ \text{s.t.} \quad & \gamma_n = \sum_{m=1}^M \rho_{k,n,m} \psi_{n,m}, \quad \forall n = 1, \dots, N, \end{aligned} \quad (11)$$

where $\boldsymbol{\gamma} := (\gamma_1, \dots, \gamma_N)^T$ and

$$U_n(x, \mathbf{w}) := \begin{cases} u_n x - v_n |x - R_n|, & \text{for } n = 1, \dots, N_{\text{RT}} \\ (1 + u_n)x, & \text{for } n = N_{\text{RT}} + 1, \dots, N. \end{cases} \quad (12)$$

In the optimization problem (11), $\gamma_n \in \Re$ for all n and $\boldsymbol{\psi}_m \in \mathcal{A}$ for all m , where $\boldsymbol{\psi}_m := (\psi_{1,m}, \dots, \psi_{N,m})^T$ and $\mathcal{A} := \{(x_1, \dots, x_N)^T \mid 0 \leq x_n \leq 1 \text{ for all } n, \sum_{n=1}^N x_n \leq 1\}$. We define $\bar{\boldsymbol{\psi}} := (\psi_{n,m})_{\substack{n=1, \dots, N \\ m=1, \dots, M}}$. Let Ψ^* be the set of all optimal solutions corresponding to $\boldsymbol{\psi}$. Note that $\mathcal{S}_k(\mathbf{w})$ is the set of all optimal solutions corresponding to $\boldsymbol{\gamma}$. The following relationship holds for Ψ^* and $\mathcal{S}_k(\mathbf{w})$ from the constraints in (11).

$$\mathcal{S}_k(\mathbf{w}) := \left\{ \left(\sum_{m=1}^M \rho_{k,1,m} \psi_{1,m}, \dots, \sum_{m=1}^M \rho_{k,N,m} \psi_{N,m} \right)^T \mid \bar{\boldsymbol{\psi}} \in \Psi^* \right\}. \quad (13)$$

We apply the dual optimization technique to solve the optimization problem (11). The Lagrangian is as follows.

$$\begin{aligned} \Upsilon(\boldsymbol{\gamma}, \bar{\boldsymbol{\psi}}, \boldsymbol{\lambda}) := & \sum_{n=1}^N \{U_n(\gamma_n, \mathbf{w}) - \lambda_n \gamma_n\} \\ & + \sum_{m=1}^M \sum_{n=1}^N \lambda_n \rho_{k,n,m} \psi_{n,m} \end{aligned} \quad (14)$$

where λ_n 's are Lagrange multipliers and $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_N)^T$.

The dual function is as follows:

$$\begin{aligned} f(\boldsymbol{\lambda}) := & \sum_{n=1}^N \max_{\gamma_n \in \Re} \{U_n(\gamma_n, \mathbf{w}) - \lambda_n \gamma_n\} \\ & + \sum_{m=1}^M \max_{\boldsymbol{\psi}_m \in \mathcal{A}} \sum_{n=1}^N \lambda_n \rho_{k,n,m} \psi_{n,m}. \end{aligned} \quad (15)$$

From (12), we have $f(\boldsymbol{\lambda}) < \infty$ if and only if $u_n - v_n \leq \lambda_n \leq u_n + v_n$ for $n = 1, \dots, N_{\text{RT}}$ and $\lambda_n = 1 + u_n$ for $n = N_{\text{RT}} + 1, \dots, N$. Therefore, we have the following dual problem:

$$\begin{aligned} \min \quad & f(\boldsymbol{\lambda}) \\ \text{s.t.} \quad & u_n - v_n \leq \lambda_n \leq u_n + v_n, \quad \forall n = 1, \dots, N_{\text{RT}}, \\ & \lambda_n = 1 + u_n, \quad \forall n = N_{\text{RT}} + 1, \dots, N. \end{aligned} \quad (16)$$

Let Λ^* be the set of the solutions of (16).

Let us define

$$\mathcal{E}_m(\boldsymbol{\lambda}) := \arg \max_{\mathbf{x} \in \mathcal{A}} \sum_{n=1}^N \lambda_n \rho_{k,n,m} x_n \quad (17)$$

where $\mathbf{x} := (x_1, \dots, x_N)^T$. We have $\mathbf{x} \in \mathcal{E}_m$ for \mathbf{x} such that $x_y = 1$ and $x_n = 0$ for other n 's, where $y \in \arg \max_{n=1, \dots, N} \lambda_n \rho_{k,n,m}$.

We also define $\mathcal{F}(\boldsymbol{\lambda})$ as follows:

$$\mathcal{F}(\boldsymbol{\lambda}) := \left\{ \left(\sum_{m=1}^M \rho_{k,1,m} \psi_{1,m}, \dots, \sum_{m=1}^M \rho_{k,N,m} \psi_{N,m} \right)^T \mid \boldsymbol{\psi}_m \in \mathcal{E}_m(\boldsymbol{\lambda}), \text{ for all } m = 1, \dots, M \right\}. \quad (18)$$

The set $\mathcal{F}(\boldsymbol{\lambda})$ can also be expressed as $\mathcal{F}(\boldsymbol{\lambda}) = \arg \max_{\mathbf{x} \in \mathcal{C}_k} \boldsymbol{\lambda}^T \mathbf{x}$. Since the optimization problem (11) is convex and strictly feasible, the strong duality holds from Slater's constraint qualification [13, p. 520]. Therefore, we have $\Upsilon(\boldsymbol{\gamma}, \bar{\boldsymbol{\psi}}, \boldsymbol{\lambda}^*) = f(\boldsymbol{\lambda}^*)$ for $\boldsymbol{\gamma} \in \mathcal{S}_k(\mathbf{w})$, $\bar{\boldsymbol{\psi}} \in \Psi^*$, and $\boldsymbol{\lambda}^* \in \Lambda^*$. Hence, we have $\boldsymbol{\psi}_m \in \mathcal{E}_m(\boldsymbol{\lambda}^*)$ for all m , $\bar{\boldsymbol{\psi}} \in \Psi^*$, and $\boldsymbol{\lambda}^* \in \Lambda^*$. From (13) and (18), we can conclude that $\mathcal{S}_k(\mathbf{w}) \subset \mathcal{F}(\boldsymbol{\lambda}^*)$. Since we can calculate a vector in $\mathcal{F}(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}$, it is possible to find $\mathbf{s}_k(\mathbf{w})$ if we derive $\boldsymbol{\lambda}^*$.

Therefore, we now derive the dual optimal solution, $\boldsymbol{\lambda}^*$. Let $\partial f(\boldsymbol{\lambda})$ denote the subdifferential (i.e., the set of all subgradients) of f at $\boldsymbol{\lambda}$. We can calculate $\partial f(\boldsymbol{\lambda})$ as follows [13, p. 604]:

$$\partial f(\boldsymbol{\lambda}) := \{\mathbf{x} - \mathbf{y} \mid \mathbf{x} \in \mathcal{F}(\boldsymbol{\lambda}) \text{ and } y_n \in \mathcal{G}_n(\lambda_n), \text{ for all } n = 1, \dots, N\} \quad (19)$$

where $\mathbf{x} := (x_1, \dots, x_N)^T$, $\mathbf{y} := (y_1, \dots, y_N)^T$,

$$\begin{aligned} \mathcal{G}_n(\lambda_n) := & \arg \max_x \{U_n(x, \mathbf{w}) - \lambda_n x\} \\ = & \begin{cases} \{x \mid x \geq R_n\}, & \text{if } \lambda_n = u_n - v_n \\ \{R_n\}, & \text{if } u_n - v_n < \lambda_n < u_n + v_n \\ \{x \mid x \leq R_n\}, & \text{if } \lambda_n = u_n + v_n \end{cases} \end{aligned} \quad (20)$$

for $n = 1, \dots, N_{\text{RT}}$, and $\mathcal{G}_n(\lambda_n) = \Re$ for $n = N_{\text{RT}} + 1, \dots, N$. In addition, we define $\mathcal{V}_n(\boldsymbol{\lambda})$ as the subdifferential of $f(\boldsymbol{\lambda})$ with respect to λ_n . Then, we have $\mathcal{V}_n(\boldsymbol{\lambda}) := \{x_n \mid \mathbf{x} \in \partial f(\boldsymbol{\lambda})\}$. Since f is a convex function, we have $\boldsymbol{\lambda} \in \Lambda^*$ if $0 \in \mathcal{V}_n(\boldsymbol{\lambda})$ for all n .

Let $\theta_n(\boldsymbol{\lambda})$ be the minimum value among x 's such that $0 \in \mathcal{V}_n((\lambda_1, \dots, \lambda_{n-1}, x, \lambda_{n+1}, \dots, \lambda_N)^T)$. We also define $\boldsymbol{\theta}(\boldsymbol{\lambda}) := (\theta_1(\boldsymbol{\lambda}), \dots, \theta_N(\boldsymbol{\lambda}))^T$. Then, we have $\boldsymbol{\lambda} \in \Lambda^*$ for $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda} = \boldsymbol{\theta}(\boldsymbol{\lambda})$. Algorithm 1 finds such $\boldsymbol{\lambda}$ by iteratively updating $\boldsymbol{\lambda}^{(i)} = (\lambda_1^{(i)}, \dots, \lambda_N^{(i)})^T$, i.e., an estimation of $\boldsymbol{\lambda}^*$ at the i th iteration. In the beginning, the algorithm sets $\lambda_n^{(0)}$ to the smallest possible value for all n . At each iteration, the algorithm selects each user in turn and adjusts the estimation for that user. Specifically, at the user j 's turn of the i th iteration, the algorithm updates $\lambda_j^{(i)}$ to $\theta_j(\boldsymbol{\lambda}^{(i)})$. This eventually leads to the convergence of $\boldsymbol{\lambda}^{(i)}$ to $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda} = \boldsymbol{\theta}(\boldsymbol{\lambda})$, i.e., $\boldsymbol{\lambda}^*$. Algorithm 1 also updates $\mathbf{s}^{(i)} = (s_1^{(i)}, \dots, s_N^{(i)})^T$, which is the estimation of $\mathbf{s}_k(\mathbf{w})$ at the i th iteration. The algorithm updates $\mathbf{s}^{(i)}$ every iteration, and it is satisfied that $\mathbf{s}^{(i)} \in \mathcal{F}(\boldsymbol{\lambda}^{(i)})$. Therefore, $s_j^{(i)}$ is also used to judge whether $0 \in \mathcal{V}_j(\boldsymbol{\lambda}^{(i)})$ is satisfied (i.e., $\lambda_j^{(i)} = \theta_j(\boldsymbol{\lambda}^{(i)})$ is satisfied) at

the user j 's turn. In Algorithm 1, we use a_m to represent subcarrier allocation instead of $\psi_{n,m}$. The variable a_m is defined as the index of the user to which subcarrier m is allocated. If $a_m = j$, then $\psi_{j,m} = 1$ and $\psi_{n,m} = 0$ for all $n \neq j$. On the basis of a_m , the algorithm calculates $\mathbf{s}^{(i)}$. For $\mathbf{s}^{(i)}$ to satisfy $\mathbf{s}^{(i)} \in \mathcal{F}(\boldsymbol{\lambda}^{(i)})$, the variable a_m always satisfies the condition $a_m \in \arg \max_{n=1,\dots,N} \lambda_n^{(i)} \rho_{k,n,m}$.

We now explain the detailed operation of Algorithm 1. In lines 2 and 3 of Algorithm 1, $\lambda_n^{(0)}$'s are set to be the smallest possible values. In lines 4–8, the algorithm calculates $\mathbf{s}^{(0)}$ satisfying $\mathbf{s}^{(0)} \in \mathcal{F}(\boldsymbol{\lambda}^{(0)})$. Iterations begin at line 9. The number of iterations is denoted by I . At the i th iteration, the algorithm selects each RT user in sequence (line 12). At the user j 's turn, the algorithm increases the value of $\lambda_j^{(i)}$ until it is satisfied that $\lambda_j^{(i)} = \theta_j(\boldsymbol{\lambda}^{(i)})$.

In line 13, \mathcal{Y} is the set of the subcarriers that are not allocated to the user j . In line 15, among \mathcal{Y} , the algorithm chooses a subcarrier that can be reallocated to the user j by increasing $\lambda_j^{(i)}$ in the smallest amount, and assigns the index of the subcarrier to m^* . In line 17, the algorithm calculates λ_{tmp} that is the required $\lambda_j^{(i)}$ for reallocating the subcarrier m^* . If λ_{tmp} exceeds the highest possible value of $\lambda_j^{(i)}$, i.e., $u_j + v_j$, the algorithm sets $\lambda_j^{(i)}$ to $u_j + v_j$ (lines 18 and 19). Otherwise, the algorithm increases $\lambda_j^{(i)}$ to λ_{tmp} , reallocates the subcarrier m^* , recalculates $\mathbf{s}^{(i)}$, and removes the subcarrier m^* from \mathcal{Y} (lines 20–25). For the user j , this subcarrier reallocation procedure continues until $\lambda_j^{(i)} = \theta_j(\boldsymbol{\lambda}^{(i)})$ is satisfied. If $\lambda_j^{(i)} = u_j + v_j$, we have $\{x | x \geq s_j^{(i)} - R_j\} \subset \mathcal{V}_j(\boldsymbol{\lambda}^{(i)})$. If $u_j - v_j < \lambda_j^{(i)} < u_j + v_j$, we have $\max\{\mathcal{V}_j(\boldsymbol{\lambda}^{(i)})\} = s_j^{(i)} - R_j$. Therefore, the subcarrier reallocation procedure stops when $\lambda_j^{(i)}$ reaches $u_j + v_j$ or $s_j^{(i)}$ exceeds R_j (line 14). If \mathcal{Y} becomes empty, the algorithm stops subcarrier reallocation (line 14) and sets $\lambda_j^{(i)}$ to $u_j + v_j$ (lines 28–30). In this case, we also have $\lambda_j^{(i)} = \theta_j(\boldsymbol{\lambda}^{(i)})$.

On the assumption that I is infinite, the following theorem states that the algorithm finds $\boldsymbol{\lambda}^* \in \boldsymbol{\Lambda}^*$.

Theorem 1: As $i \rightarrow \infty$, $\boldsymbol{\lambda}^{(i)}$ converges to $\boldsymbol{\lambda}^* \in \boldsymbol{\Lambda}^*$.

Proof: See Appendix A. ■

Algorithm 1: Calculating $\boldsymbol{\lambda}^*$ and $\mathbf{s}_k(\mathbf{w})$

1 **begin**

2 $\lambda_n^{(0)} \leftarrow u_n - v_n$ for $n = 1, \dots, N_{\text{RT}}$;

3 $\lambda_n^{(0)} \leftarrow 1 + u_n$ for $n = N_{\text{RT}} + 1, \dots, N$;

4 $s_n^{(0)} \leftarrow 0$ for all n ;

5 **for** $m = 1$ **to** $m = M$ **do**

6 $a_m \leftarrow \arg \max_{n=1,\dots,N} \lambda_n^{(0)} \rho_{k,n,m}$;

7 $s_{a_m}^{(0)} \leftarrow s_{a_m}^{(0)} + \rho_{k,a_m,m}$;

8 **end**

9 **for** $i = 1$ **to** $i = I$ **do**

10 $\lambda_n^{(i)} \leftarrow \lambda_n^{(i-1)}$ for all n ;

11 $s_n^{(i)} \leftarrow s_n^{(i-1)}$ for all n ;

12 **for** $j = 1$ **to** $j = N_{\text{RT}}$ **do**

13 $\mathcal{Y} \leftarrow \{m | a_m \neq j\}$;

14 **while** $\mathcal{Y} \neq \{\}$ and $\lambda_j^{(i)} < u_j + v_j$ and $s_j^{(i)} < R_j$ **do**

15 $m^* \leftarrow \arg \min_{m \in \mathcal{Y}} \lambda_{a_m}^{(i)} \rho_{k,a_m,m} / \rho_{k,j,m}$;

16 $n^* \leftarrow a_{m^*}$;

17 $\lambda_{\text{tmp}} \leftarrow \lambda_{n^*}^{(i)} \rho_{k,n^*,m^*} / \rho_{k,j,m^*}$;

18 **if** $\lambda_{\text{tmp}} > u_j + v_j$ **then**

19 $\lambda_j^{(i)} \leftarrow u_j + v_j$;

20 **else**

21 $\lambda_j^{(i)} \leftarrow \lambda_{\text{tmp}}$;

22 $a_{m^*} \leftarrow j$;

23 $s_{n^*}^{(i)} \leftarrow s_{n^*}^{(i)} - \rho_{k,n^*,m^*}$;

24 $s_j^{(i)} \leftarrow s_j^{(i)} + \rho_{k,j,m^*}$;

25 $\mathcal{Y} \leftarrow \mathcal{Y} - \{m^*\}$;

26 **end**

27 **end**

28 **if** $\mathcal{Y} = \{\}$ and $\lambda_j^{(i)} < u_j + v_j$ and $s_j^{(i)} < R_j$ **then**

29 $\lambda_j^{(i)} \leftarrow u_j + v_j$;

30 **end**

31 **end**

32 **end**

33 **end**

Practically, $\boldsymbol{\lambda}^{(i)}$ converges very fast, therefore, a solution very close to the optimal one can be found with only a few iterations.

Generally, OFDMA wireless systems use a large number (e.g., hundreds or thousands) of subcarriers, and the subcarriers have different SNRs due to frequency selective fading. In this case, we can assume that $|\mathcal{F}(\boldsymbol{\lambda})| = 1$ for all $\boldsymbol{\lambda}$ since the number of subcarriers such that $|\arg \max_{n=1,\dots,N} \lambda_n \rho_{k,n,m}| > 1$ (for subcarrier m) is very small compared to the total number of subcarriers, M . We will assume this for the rest of the paper. Since $\mathbf{s}_k(\mathbf{w}) \in \mathcal{S}_k(\mathbf{w}) \subset \mathcal{F}(\boldsymbol{\Lambda}^*)$, we have $|\mathcal{S}_k(\mathbf{w})| = 1$ and $\mathcal{F}(\boldsymbol{\Lambda}^*) = \{\mathbf{s}_k(\mathbf{w})\}$ from this assumption. In addition, we have $\mathcal{F}(\boldsymbol{\lambda}^{(i)}) = \{\mathbf{s}^{(i)}\}$. Since $\mathcal{F}(\boldsymbol{\lambda})$ is a continuous mapping, we can conclude that $\mathbf{s}^{(i)} \rightarrow \mathbf{s}_k(\mathbf{w})$ as $i \rightarrow \infty$.

D. Calculation of Optimal Solutions, \mathbf{w}^* and $\bar{\mathbf{r}}^*$

In Section III-C, we have suggested Algorithm 1 which finds the solution to (11) on a frame-by-frame basis. Now, we use the projection stochastic subgradient method [14] to find the solution to the problem (9), denoted by $\mathbf{w}^* (\in \mathcal{W}^*)$. We define

$$\mathbf{w}(t) := (u_1(t), \dots, u_N(t), v_1(t), \dots, v_{N_{\text{RT}}}(t))^T \quad (21)$$

as the estimation of \mathbf{w}^* at frame t . The projection stochastic subgradient method updates $\mathbf{w}(t)$ iteratively, and $\mathbf{w}(t)$ converges to

\mathbf{w}^* . Without loss of generality, we assume that the iteration begins at frame 1. The initial value is $\mathbf{w}(1)$ such that $\mathbf{w}(1) \succeq \mathbf{0}$, and the iteration at frame t is as follows:

$$\mathbf{w}(t+1) = \Pi[\mathbf{w}(t) - \delta(t)\boldsymbol{\xi}(t)] \quad (22)$$

where $\Pi[\mathbf{x}] = (\max\{0, x_1\}, \dots, \max\{0, x_N\})^T$ for $\mathbf{x} = (x_1, \dots, x_N)^T$.

Let $\mathbf{s}(t, \mathbf{w}) = (s_1(t, \mathbf{w}), \dots, s_N(t, \mathbf{w}))^T$ be a random vector that satisfies $\mathbf{s}(t, \mathbf{w}) = \mathbf{s}_k(\mathbf{w})$ when the channel state at the frame t is k . In (22), $\boldsymbol{\xi}(t)$ is defined as

$$\boldsymbol{\xi}(t) := (\xi_1(t), \dots, \xi_N(t), \zeta_1(t), \dots, \zeta_{N_{\text{RT}}}(t))^T \quad (23)$$

where

$$\xi_n(t) := s_n(t, \mathbf{w}(t)) - R_n \quad (24)$$

for $n = 1, \dots, N$ and

$$\zeta_n(t) := D_n - |s_n(t, \mathbf{w}(t)) - R_n| \quad (25)$$

for $n = 1, \dots, N_{\text{RT}}$, and $\delta(t)$ is the step size that satisfies the following conditions:

$$\delta(t) > 0, \quad \sum_{t=1}^{\infty} \delta(t) = \infty, \quad \sum_{t=1}^{\infty} \delta(t)^2 < \infty. \quad (26)$$

For example, $\delta(t) = c/t$ where c is a positive constant.

The following theorem states that the sequence $\{\mathbf{w}(t)\}$ has a limit in \mathcal{W}^* .

Theorem 2: If it is assumed that the channel state at a frame is independent of the channel states at the previous frames, $\{\mathbf{w}(t)\}$ has a limit in \mathcal{W}^* with probability 1.

Proof: See Appendix B. ■

We will discuss the assumption in Theorem 2 later in this section.

We assume that the optimization problem (6) is strictly feasible. That is, there exists $\bar{\mathbf{r}}$ in the relative interior of $\bar{\mathcal{C}}$, which satisfies $\sum_{k \in \mathcal{K}} p_k r_{k,n} > R_n$ for $n = 1, \dots, N$ and $\sum_{k \in \mathcal{K}} p_k |r_{k,n} - R_n| < D_n$ for $n = 1, \dots, N_{\text{RT}}$. This assumption is trivial since it is almost the same as the feasibility condition. Since the problem (6) is convex and we assume that it is strictly feasible, the strong duality holds from Slater's constraint qualification. Therefore, we have $L(\bar{\mathbf{r}}^*, \mathbf{w}^*) = g(\mathbf{w}^*)$ for $\mathbf{w}^* \in \mathcal{W}^*$. This means that $\mathbf{r}_k^* \in \mathcal{S}_k(\mathbf{w}^*)$ and $\mathcal{S}_k(\mathbf{w}^*) = \{\mathbf{r}_k^*\}$ for all k since $|\mathcal{S}_k(\mathbf{w})| = 1$ for all \mathbf{w} . We take $\mathbf{s}_k(\mathbf{w}(t))$ as the estimation of the primal solution, \mathbf{r}_k^* , at frame t . We have $\mathcal{S}_k(\mathbf{w}(t)) = \{\mathbf{s}_k(\mathbf{w}(t))\}$. Since $\{\mathbf{w}(t)\}$ has a limit in \mathcal{W}^* with probability 1 and $\mathcal{S}_k(\mathbf{w})$ is a continuous mapping, we can conclude that $\mathbf{s}_k(\mathbf{w}(t)) \rightarrow \mathbf{r}_k^*$ with probability 1. Even when $\mathbf{w}(t)$ is not converged sufficiently, $\mathbf{s}(t, \mathbf{w}(t))$ can be used as a good estimation of the optimal transmission rate vector at frame t . Therefore, we adopt $s_n(t, \mathbf{w}(t))$ as the transmission rate of user n at frame t .

Remark 1: The assumption in Theorem 2 is too strong since the channel state generally depends on the previous channel states. Fortunately, $\mathbf{w}(t)$ converges well to the dual solution without the assumption of the independent channel state. We will show it by the simulation in Section IV.

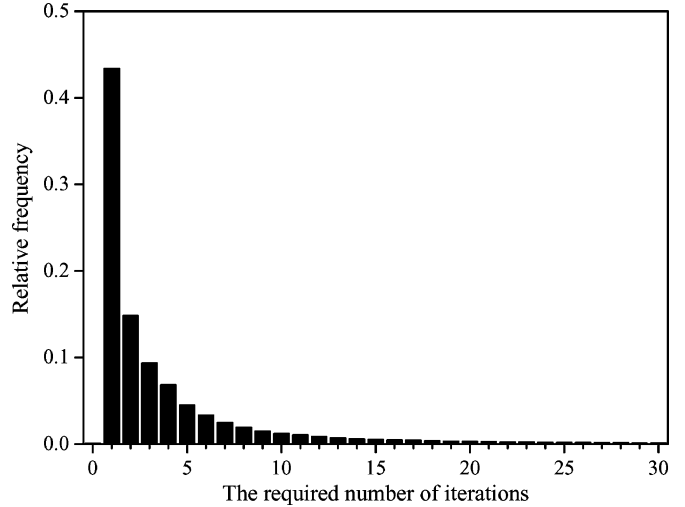


Fig. 1. The distribution of the required number of iterations for convergence in Algorithm 1.

Remark 2: Instead of the step size satisfying (26), we will use the constant step size, $\delta(t) = c$, where c is a positive constant. If the constant step size is used, although $\mathbf{w}(t)$ does not converge precisely, the projection stochastic subgradient method can continually adapt to the non-stationary channel condition and the varying constraints. Therefore, the constant step size is more appropriate in practical systems than that satisfying (26). We will also show by the simulation in Section IV that when the constant step size is used, $\mathbf{w}(t)$ nearly converges to the dual solution and the primal solution can be approximately derived.

IV. SIMULATION RESULTS

We conducted computer simulation to show the practical validity of the assumptions and approximations used in the previous sections and to demonstrate the performance of the proposed algorithm.

In simulation, the frame duration is 4 ms and one frame contains 10 OFDM symbols of which the duration is 0.4 ms. The carrier frequency is 2 GHz. There are 512 subcarriers which are spaced by 2.5 kHz. The cell is circular and its radius is 1 km. The moving speed of users is 50 km/h invariably. If a user steps over the cell boundary, it is relocated to the opposite side of the cell.

We examine the stationary and non-stationary channel conditions. For the stationary channel condition, the multipath fading is only considered. The multipath fading process is generated by the wide sense stationary uncorrelated scattering (WSSUS) channel model [18] with the exponentially decaying power delay profile of which the average delay spread is 1 μ s. For the non-stationary channel condition, the path loss and the shadowing are also included in computation of a channel gain. The path loss is calculated as (path loss) = $15.3 + 37.6 \log_{10} d$, where d is the distance (in meters) between the BS and the user. The log-normal shadowing model with the zero mean and the standard deviation of 8 dB is used. We assume that the total available power of the BS is 37 dBm and the noise density (i.e., $N_o/2$) is -164 dBm/Hz.

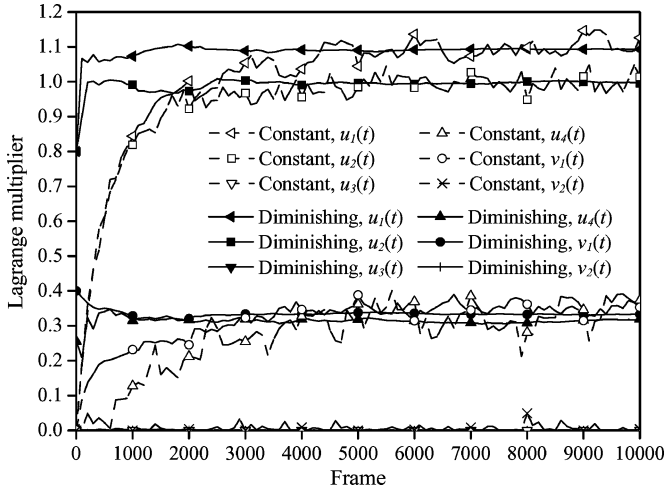


Fig. 2. Convergence of Lagrange multipliers.

The traffic generation rate of RT traffic is 512 kb/s. At each frame, an RT traffic source produces a packet with the fixed size of 256 bytes. We define T_n as the time within which the RT packets of user n should be delivered to the user after arriving. The RT packets of user n are dropped at the BS when T_n elapses after arriving. The packet drop rate which is the proportion of the dropped packets to the total generated packets is used as a performance metric for the RT users. The BE users are assumed to have an infinite backlog.

Fig. 1 shows the distribution of the required number of iterations for convergence in Algorithm 1. For the simulation, we do not limit the maximum number of iterations (i.e., I), but stop iteration once λ^i converges to λ^* . We have run the simulation for 100,000 frames, counted the required number of iterations every frame, and drawn its distribution. The stationary channel condition is used for the simulation. There are four RT users and four BE users, of which the parameters are $R_1 = R_2 = R_3 = R_4 = 200$ kb/s, $R_5 = R_6 = R_7 = R_8 = 0$ kb/s, and $D_1 = D_2 = D_3 = D_4 = 100$ kb/s. We can see that λ^i converges completely within 20 iterations at almost all frames (exactly, 95 percent of frames). Even in the frames where the complete convergence takes more than 20 iterations, λ^i converges very close to λ^* within 20 iterations. Thus, we will set I as 20 for the rest of simulations.

Fig. 2 shows the convergence of the Lagrange multipliers in the system with two RT users and two BE users. The stationary channel condition is used to obtain the results of Fig. 2. For Fig. 2, we apply both the diminishing step size of $\delta(t) = 0.002/t$ and the constant step size of $\delta(t) = 0.000005$. When the WSSUS channel model is used, the channel state at a frame is dependent on the channel states at the previous frames. However, in the simulation using the diminishing step size, to realize the assumption in Theorem 2, we apply the independent channel that is made by randomly rearranging the generated channel states. Then, the simulation using the diminishing step size fully complies with the condition of the independent channel states for Theorem 2. For the simulation using the constant step size, we use the WSSUS channel model. We set R_n and D_n as follows: $R_1 = 400$ kb/s, $R_2 = 400$ kb/s, $R_3 = 0$ kb/s, $R_4 = 700$ kb/s, $D_1 = 200$ kb/s, and $D_2 = 400$ kb/s.

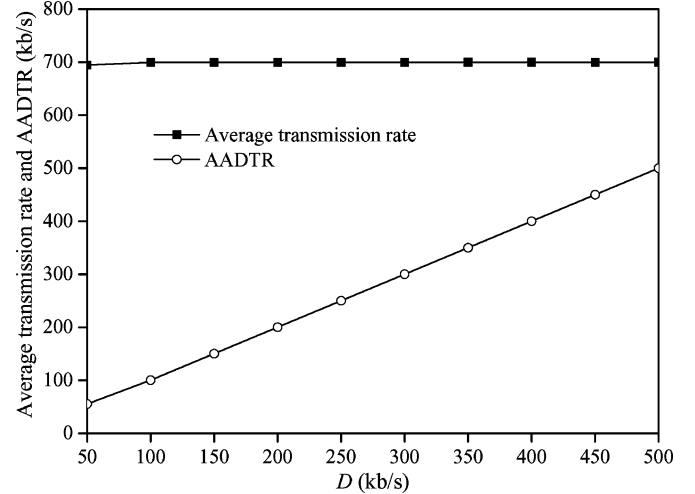


Fig. 3. The average transmission rate and AADTR of RT users according to D .

Fig. 2 shows that around $t = 10000$, the Lagrange multipliers for the diminishing step size respectively converge to $u_1(t) = 1.09$, $u_2(t) = 0.99$, $u_3(t) = 0$, $u_4(t) = 0.32$, $v_1(t) = 0.33$, and $v_2(t) = 0$. It is also seen that the Lagrange multipliers of the constant step size fluctuate around those of the diminishing step size. On the other hand, we have obtained the average transmission rate and AADTR by averaging over 100 000 frames. For the diminishing step size, the average transmission rates of the users 1–4 and AADTRs of the users 1 and 2 are respectively 400, 401, 391, 700, 201, and 362 kb/s. For the constant step size, they are 398, 398, 394, 700, 201, and 359 kb/s. Considering that the QoS requirements are given as $R_1 = 400$, $R_2 = 400$, $R_3 = 0$, $R_4 = 700$, $D_1 = 200$, and $D_2 = 400$ kb/s, we see that these are well satisfied. In addition, these results are almost the same for both step size rules. Therefore, it can be concluded that the proposed algorithm with the constant step size performs well without the assumption of the independent channel states. We will use the constant step size, $\delta(t) = 0.000005$, for the rest of simulations.

Figs. 3–5 show the performance of the proposed algorithm under the non-stationary channel condition. The simulation time is 3 000 000 frames. There are four RT users and four BE users. The QoS requirements are as follows: $R_1 = R_2 = R_3 = R_4 = 700$ kb/s, $R_5 = 0$ kb/s, $R_6 = 500$ kb/s, $R_7 = 1000$ kb/s, and $R_8 = 1500$ kb/s. The graphs are plotted as a function of $D (= D_1 = D_2 = D_3 = D_4)$.

Fig. 3 shows the average transmission rate and AADTR of the RT users which are averaged over the whole simulation time and all RT users. The average transmission rate is about 700 kb/s for all range of D , and AADTR is almost the same value as D .

Fig. 4 depicts the packet drop rates of RT users, when $T_1 = 200$ ms, $T_2 = 500$ ms, $T_3 = 1000$ ms, and $T_4 = 2000$ ms. This figure shows that the packet drop rate can be reduced by decreasing the value of D . In this figure, we can see that the packet drop rate is a function of both D_n and T_n . Therefore, by simulations or field trials, it is possible to find the required D_n to achieve a certain packet drop rate when T_n is given. It can be used to decide the tolerable AADTR when an RT connection is requested.

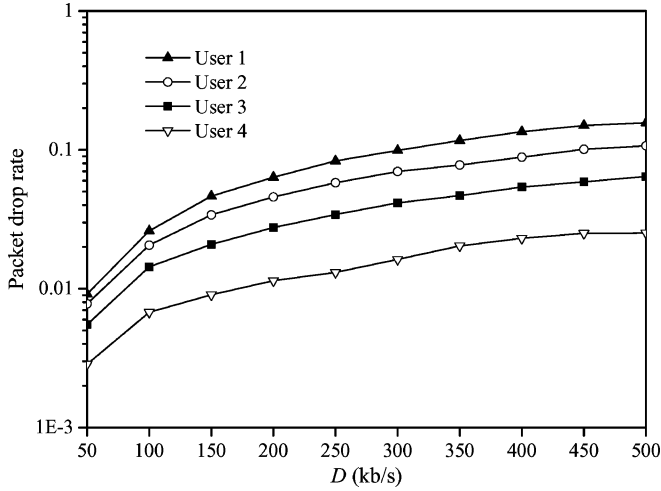
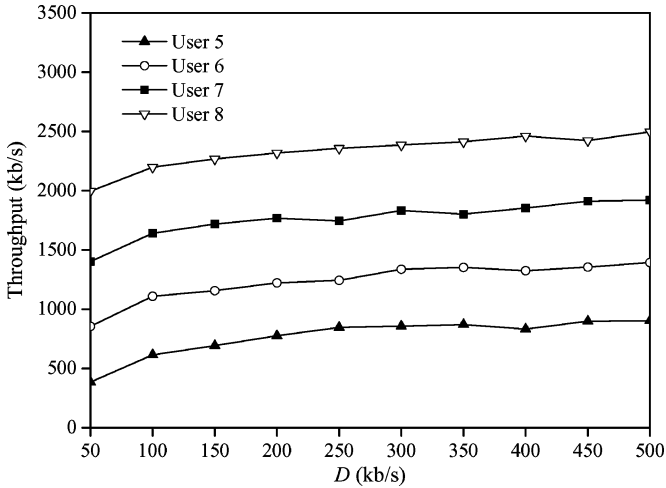

 Fig. 4. The packet drop rate of RT users according to D .

 Fig. 5. The throughput of BE users according to D .

Fig. 5 shows the average throughput of each of four BE users. It is noted that the required average transmission rates of these BE users are given as $R_5 = 0$ kb/s, $R_6 = 500$ kb/s, $R_7 = 1000$ kb/s, and $R_8 = 1500$ kb/s. The throughput is higher for the user with higher required average transmission rate, since even when the channel condition is generally bad (e.g., when the user is far from the BS), the required average transmission rate is guaranteed with the proposed algorithm. On the other hand, we can see from the figure that the throughputs of BE users are lower with the smaller D . This is because, as D decreases, the drop rate of RT packets is reduced and the less resource is allocated to the BE users.

In Figs. 6 and 7, we compare the proposed resource allocation algorithm with the modified largest weighted delay first (M-LWDF) [8] in the packet drop rate of the RT users and the throughput of the BE users. M-LWDF is chosen for the comparison since it supports the similar QoS requirements to the proposed algorithm. M-LWDF supports the RT and BE services simultaneously, and aims to reduce the packet drop rate of the RT users and guarantee the required average transmission rate of the BE users. Since M-LWDF is originally designed for TDMA systems, it selects the user who is served every frame. We modify M-LWDF for OFDMA so as to select the served user for each subcarrier every frame as follows.

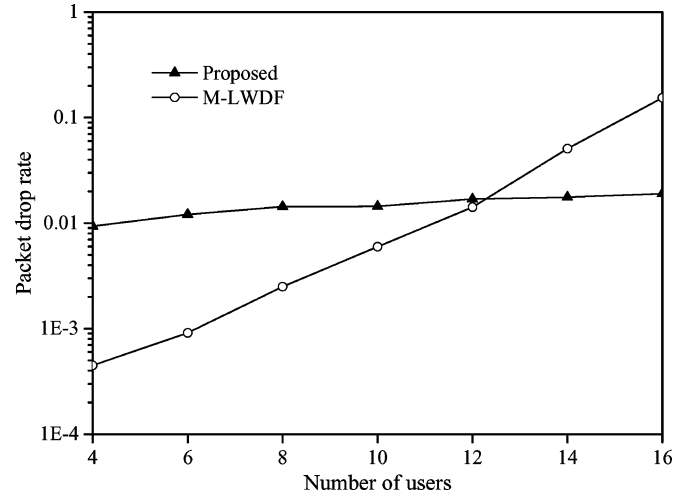


Fig. 6. The packet drop rates of the proposed algorithm and M-LWDF according to the number of users.

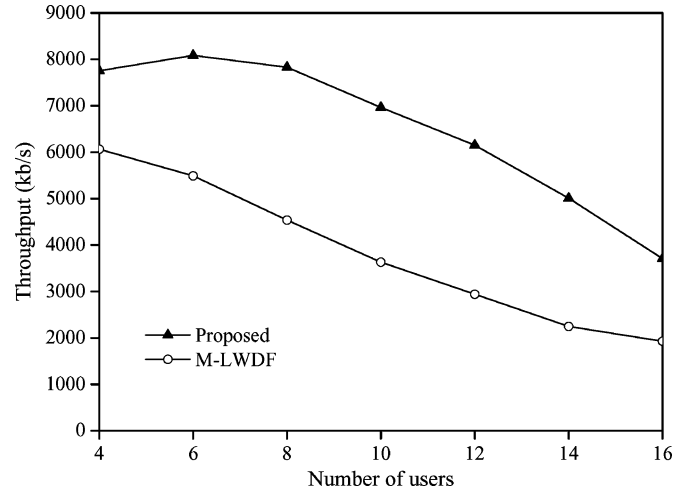


Fig. 7. The throughputs of the proposed algorithm and M-LWDF according to the number of users.

M-LWDF makes the scheduling decision on the basis of the current channel states and the transmission queue states of users. In the simulation herein, M-LWDF serves the user n for which $a_n \omega_n \rho_{n,m}(t)$ is maximized for the subcarrier m at frame t , where a_n and ω_n are set in different ways according to the class of user n . For the RT user n , ω_n is the head-of-the-line (HOL) packet delay of user n , and $a_n = 1/(T_d \rho_{n,m}^{\text{avg}}(t))$, where $\rho_{n,m}^{\text{avg}}(t)$ is an average of $\rho_{n,m}(t)$ and is calculated as

$$\rho_{n,m}^{\text{avg}}(t) = \begin{cases} 0.001 \rho_{n,m}(t) + 0.999 \rho_{n,m}^{\text{avg}}(t-1), & \text{if } m = 1 \\ 0.001 \rho_{n,m}(t) + 0.999 \rho_{n,m-1}^{\text{avg}}(t), & \text{otherwise.} \end{cases} \quad (27)$$

For the BE user n , there is a virtual token bucket where tokens arrive at the minimum average transmission rate, R_n , and are reduced by the actual amount of data served. If the number of bits in the token bucket of the BE user n is denoted by x_n , then $\omega_n = x_n/R_n$. The value a_n for the BE users should be decided to balance the priorities of the RT and BE users. We set $a_n = 0.1$ for the BE users. For more detailed operation of M-LWDF, refer to [8].

Figs. 6 and 7 respectively plot the packet drop rate of RT users and the total throughput of BE users according to the number of users. There are the same number of the RT and BE users.

For the RT packets, $T_n = 2000$ ms for $n = 1, \dots, N_{\text{RT}}$. The simulation time is 3 000 000 frames. The QoS requirements for the proposed algorithm are given as $R_n = 700$ kb/s for $n = 1, \dots, N_{\text{RT}}$, $R_n = 200$ kb/s for $n = N_{\text{RT}} + 1, \dots, N$, and $D_n = 300$ kb/s for $n = 1, \dots, N_{\text{RT}}$. For M-LWDF, $R_n = 200$ kb/s for $n = N_{\text{RT}} + 1, \dots, N$.

In Fig. 6, the packet drop rate of the proposed algorithm remains stable regardless of the number of users, whereas the packet drop rate of M-LWDF increases according to the number of users. It means that the proposed algorithm is able to provide a stable QoS that is not influenced by the varying loads.

In Fig. 7, it is seen that the proposed algorithm outperforms M-LWDF in the total throughput of BE users. This is because even if M-LWDF also exploits the channel variation, it is designed only so as to guarantee the required average transmission rate of BE users, not to maximize the system throughput.

V. CONCLUSION

We have suggested the resource allocation algorithm for the OFDMA system, which accommodates both RT and BE users under the time-varying channel condition. The proposed algorithm aims to maximize the system throughput while satisfying the QoS requirements of both RT and BE users. The distinctive feature of the proposed algorithm is the restriction on AADTR, which is introduced to provide stable transmission rates to the RT users.

We have formulated the optimization problem, and developed the algorithm that solves it by the dual optimization techniques. It is shown by the simulation that the proposed algorithm meets well its design goal and outperforms M-LWDF in terms of the packet drop rate of the RT users and the throughput of the BE users.

APPENDIX A

Proof of Theorem 1

To prove the theorem, we first prove the following Lemmas 1 and 2.

Lemma 1: For all $\lambda^* := (\lambda_1^*, \dots, \lambda_N^*)^T \in \Lambda^*$, we have $\lambda^* \succeq \theta(\lambda^{(i)}) \succeq \lambda^{(i)}$ for all i .

Proof: We first prove that $\lambda^* \succeq \theta(\lambda)$ for λ that satisfies $\lambda^* \succeq \lambda$. We have $0 \in \mathcal{V}_n(\lambda^*)$ for all n . Therefore, we have $\lambda^* \succeq \theta(\lambda^*)$. Since $\theta_n(\lambda)$ is a non-decreasing function of λ_x for $x \neq n$, we can prove that $\lambda^* \succeq \theta(\lambda^*) \succeq \theta(\lambda)$ for λ that satisfies $\lambda^* \succeq \lambda$.

For the proof, we define $\lambda^{(i,j)} := (\lambda_1^{(i,j)}, \dots, \lambda_N^{(i,j)})^T$ as the value of $\lambda^{(i)}$ after the user j 's turn at the i th iteration of the algorithm. Then, we have $\lambda^{(i,0)} = \lambda^{(i-1)}$ and $\lambda^{(i,N)} = \lambda^{(i)}$. We now prove that $\lambda^* \succeq \theta(\lambda^{(0)}) \succeq \lambda^{(0)}$. And we prove that $\lambda^* \succeq \theta(\lambda^{(i,j)}) \succeq \lambda^{(i,j)}$ if $\lambda^* \succeq \theta(\lambda^{(i,j-1)}) \succeq \lambda^{(i,j-1)}$. Then the lemma can be proved.

Since $\lambda_n^{(0)}$ is the smallest possible value of λ_n , we have $\theta(\lambda^{(0)}) \succeq \lambda^{(0)}$ and $\lambda^* \succeq \lambda^{(0)}$. Therefore, $\lambda^* \succeq \theta(\lambda^{(0)}) \succeq \lambda^{(0)}$. Suppose that $\lambda^* \succeq \theta(\lambda^{(i,j-1)}) \succeq \lambda^{(i,j-1)}$. We have $\lambda_j^{(i,j)} = \theta_j(\lambda^{(i,j-1)})$ and $\lambda_n^{(i,j)} = \lambda_n^{(i,j-1)}$ for $n \neq j$. Then, we have $\lambda^* \succeq \lambda^{(i,j)}$, therefore, $\lambda^* \succeq \theta(\lambda^{(i,j)})$. Moreover, we have $\theta_j(\lambda^{(i,j)}) = \lambda_j^{(i,j)}$ and $\theta_n(\lambda^{(i,j)}) \geq \theta_n(\lambda^{(i,j-1)}) \geq$

$\lambda_n^{(i,j-1)} = \lambda_n^{(i,j)}$ for $n \neq j$, since $\theta_n(\lambda)$ is a non-decreasing function. Hence, we have $\lambda^* \succeq \theta(\lambda^{(i,j)}) \succeq \lambda^{(i,j)}$ if $\lambda^* \succeq \theta(\lambda^{(i,j-1)}) \succeq \lambda^{(i,j-1)}$. ■

Lemma 2: $\lambda^{(i+1)} \succeq \theta(\lambda^{(i)})$.

Proof: From the proof of Lemma 1, $\lambda_n^{(i,j)}$ increases as j increases for all n . Therefore, $\theta_n(\lambda^{(i,j)})$ also increases as j increases for all n , and $\theta(\lambda^{(i,j)}) \succeq \theta(\lambda^{(i)})$ for all j . Then, we have $\lambda_j^{(i+1)} = \lambda_j^{(i,j)} = \theta_j(\lambda^{(i,j-1)}) \geq \theta_j(\lambda^{(i)})$ for all j , and the lemma is proved. ■

From Lemma 1, we can learn that $\lambda_n^{(i)}$ is a non-decreasing and bounded sequence. Therefore, $\lambda^{(i)}$ converges to a vector as $i \rightarrow \infty$. Let \mathbf{x} be the vector that $\lambda^{(i)}$ converges to. From Lemma 2, we have $\mathbf{x} \succeq \theta(\lambda^{(i)})$ for all i . Since we have $\mathbf{x} \succeq \theta(\lambda^{(i)}) \succeq \lambda^{(i)}$ from Lemma 1, $\theta(\lambda^{(i)})$ also converges to \mathbf{x} . Since $\lambda^{(i)} \rightarrow \mathbf{x}$, $\theta(\lambda^{(i)}) \rightarrow \mathbf{x}$ as $i \rightarrow \infty$, and θ is a continuous function, we can conclude that $\theta(\mathbf{x}) = \mathbf{x}$. Therefore, $\mathbf{x} \in \Lambda^*$ and the theorem is proved.

APPENDIX B

Proof of Theorem 2

In [14], it is proven that $\mathbf{w}(t)$ converges to the optimal solution by the projection stochastic subgradient method if the following condition holds.

$$E\{\xi(t)|\mathbf{w}(1), \dots, \mathbf{w}(t)\} \in \partial g(\mathbf{w}(t)) \quad (28)$$

where $\partial g(\mathbf{x})$ is the subdifferential of the function g at \mathbf{x} . We have assumed that the channel state at a frame is independent of the channel states at the previous frames. Since the channel states from frame 1 to frame $t - 1$ determine $\mathbf{w}(2), \dots, \mathbf{w}(t)$, the channel state at frame t is independent of $\mathbf{w}(1), \dots, \mathbf{w}(t)$. Since $\xi(t)$ is determined by $\mathbf{w}(t)$ and the channel state at frame t from (23)–(25),

$$E\{\xi(t)|\mathbf{w}(1), \dots, \mathbf{w}(t)\} = E\{\xi(t)|\mathbf{w}(t)\} \\ = (y_1, \dots, y_N, z_1, \dots, z_{N_{\text{RT}}})^T, \quad (29)$$

where $y_n = \sum_{k \in \mathcal{K}} p_k s_{k,n}(\mathbf{w}(t)) - R_n$ for $n = 1, \dots, N$ and $z_n = D_n - \sum_{k \in \mathcal{K}} p_k |s_{k,n}(\mathbf{w}(t)) - R_n|$ for $n = 1, \dots, N_{\text{RT}}$.

The Lagrangian $L(\bar{\mathbf{r}}, \mathbf{w}(t))$ is maximized when $\mathbf{r}_k = s_k(\mathbf{w}(t))$ for all k . Therefore, the following holds from the theory of the dual subgradient [13, p. 604]:

$$(y_1, \dots, y_N, z_1, \dots, z_{N_{\text{RT}}})^T \in \partial g(\mathbf{w}(t)). \quad (30)$$

From (29) and (30), we conclude that (28) holds.

REFERENCES

- [1] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *Proc. IEEE GLOBECOM 2000*, San Francisco, CA, Nov. 2000.
- [2] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, no. 4, pp. 362–370, Dec. 2003.
- [3] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.
- [4] Y. J. Zhang and K. B. Letaief, "Energy-efficient MAC-PHY resource management with guaranteed QoS in wireless OFDM networks," in *Proc. IEEE ICC 2005*, Seoul, Korea, May 2005.

- [5] D. Kivanc, G. Li, and H. Lui, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1150–1158, Nov. 2003.
- [6] Z. Zhang, Y. He, and E. K. P. Chong, "Opportunistic downlink scheduling for multiuser OFDM systems," in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC 2005)*, New Orleans, LA, Mar. 2005, vol. 2, pp. 1206–1212.
- [7] S. S. Jeong, D. G. Jeong, and W. S. Jeon, "Cross-layer design of packet scheduling and resource allocation in OFDMA wireless multimedia networks," in *Proc. IEEE VTC 2006—Spring*, Melbourne, Australia, May 2006, vol. 1, pp. 309–313.
- [8] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.
- [9] 1xEV: 1xEvolution IS-856 TIA/EIA Standard Airlink Overview (Revision 7.1). Qualcomm Inc., May 2001.
- [10] S. Borst and P. Whiting, "Dynamic channel-sensitive scheduling algorithms for wireless data throughput optimization," *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 569–586, May 2003.
- [11] W. S. Jeon, D. G. Jeong, and B. Kim, "Packet scheduler for mobile Internet access using high speed downlink packet access systems," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1789–1801, Sept. 2004.
- [12] K. W. Choi, D. G. Jeong, and W. S. Jeon, "Packet scheduler for mobile communications systems with time-varying capacity region," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 1034–1045, Mar. 2007.
- [13] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.
- [14] Y. Ermoliev, "Stochastic quasigradient methods," in *Numerical Techniques for Stochastic Optimization*, Y. Ermoliev and R. Wets, Eds. New York: Springer-Verlag, 1988, pp. 141–185.
- [15] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Opportunistic power scheduling for multi-server wireless systems with minimum performance constraints," in *Proc. IEEE INFOCOM 2004*, Hong Kong, China, Mar. 2004, vol. 2, pp. 1067–1077.
- [16] *Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE P802.16-REVd/D4, Mar. 2004.
- [17] I. Koffman and V. Roman, "Broadband wireless access solutions based on OFDM access in IEEE 802.16," *IEEE Commun. Mag.*, vol. 40, no. 4, pp. 96–103, Apr. 2002.
- [18] P. Hoeher, "A statistical discrete-time model for the WSSUS multipath channel," *IEEE Trans. Veh. Technol.*, vol. 41, no. 4, pp. 461–468, Nov. 1992.



Kae Won Choi received the B.S. degree in civil, urban, and geosystem engineering in 2001, and the M.S. and Ph.D. degrees in electrical engineering and computer science in 2003 and 2007, respectively, all from Seoul National University, Seoul, Korea.

He is currently with Telecommunication Business of Samsung Electronics Co., Ltd., Korea. His research interests include wireless network optimization, radio resource management, wireless mesh networks, and cognitive radio.



Wha Sook Jeon (M'90–SM'01) received the B.S., M.S., and Ph.D. degrees in computer engineering from Seoul National University, Seoul, Korea, in 1983, 1985, and 1989, respectively.

From 1989 to 1999, she was with the Department of Computer Engineering, Hansung University, Korea. In 1999, she joined the faculty at Seoul National University, Korea, where she is currently a Professor in the School of Electrical Engineering and Computer Science. Her research interests include resource management for wireless and mobile net-

works, mobile communications systems, high-speed networks, communication protocols, and network performance evaluation.

Dr. Jeon currently serves on the Editorial Board of the *Journal of Communications and Networks* (JCN).



Dong Geun Jeong (S'90–M'93–SM'99) received the B.S., M.S., and Ph.D. degrees from Seoul National University, Seoul, Korea, in 1983, 1985 and 1993, respectively.

From 1986 to 1990, he was a researcher with the R&D Center of DACOM, Korea. In 1994–1997, he was with the R&D Center of Shinsegi Telecomm Inc., Korea, where he conducted and led research on advanced cellular mobile networks. In 1997, he joined the faculty at Hankuk University of Foreign Studies, Korea, where he is currently a Professor in the School

of Electronics and Information Engineering. His research interests include resource management for wireless and mobile networks, mobile communications systems, communication protocols, and network performance evaluation.

Dr. Jeong served as the TPC Vice-Chair for the IEEE VTC 2003-Spring. From 2002 to 2007, he served on the Editorial Board of the *Journal of Communications and Networks* (JCN).