A Balanced Resource Scheduling Scheme With Adaptive Priority Thresholds for OFDMA Downlink Systems

Yao-Hsing Chung, *Student Member, IEEE*, and Chung-Ju Chang, *Fellow, IEEE*

***Abstract*—This paper proposes a new balanced resource scheduling (BRS) scheme with adaptive priority thresholds for orthogonal frequency-division multiple-access (OFDMA) downlink systems. The BRS scheme achieves an excellent balance between quality-of-service (QoS) requirement guarantee and system throughput enhancement, whereas conventional schemes cannot explicitly and accurately control this tradeoff. Based on the adaptive priority threshold of each user, the BRS scheme first performs a priority-based resource allocation (RA) algorithm for users whose priority value is larger than its priority threshold to fulfill the QoS requirement. The BRS scheme then performs a channel-state-information (CSI)-based RA algorithm for the remaining users to enhance system throughput. To achieve balance between QoS guarantee and throughput enhancement, a fuzzy inference priority threshold generator adaptively and intelligently adjusts the priority threshold of each user. Simulation results show that the proposed BRS scheme with adaptive priority threshold enhances the system throughput by 16%, 8.5%, 8.2%, and 46.8% at traffic load of 0.93, compared with conventional adaptive radio RA (RRA), utility-based RRA, utility-based throughput maximization and complexity reduction scheduling, and fairness and QoS guarantee scheduling with fuzzy controls schemes, respectively, under a QoS requirement guarantee. This approach also outperforms the BRS scheme with fixed priority thresholds in both throughput enhancement and QoS guarantee.**

***Index Terms*—Fuzzy inference system, orthogonal frequency-division multiple-access (OFDMA) downlink system, quality of service (QoS), radio resource allocation (RRA), resource scheduling scheme.**

I. Introduction

ORTHOGONAL frequency-division multiplexing (OFDM) has been adopted by recent wireless communication systems as their air interface [1], [2]. In multiuser OFDM [known as orthogonal frequency-division multiple-access (OFDMA)] systems, every user may have a different channel condition on the same subchannel, creating multiuser diversity. Therefore, a resource scheduling scheme must determine which user should have access to each subchannel to fully utilize the multiuser diversity and maximize the spectrum efficiency.

Researchers have proposed many resource scheduling schemes designed for multiuser OFDM systems. Wong *et al.* [3] proposed an optimal power allocation algorithm for OFDM systems with a minimum rate constraint. Jang and Lee [4] showed that the throughput of an OFDM system can be maximized if every subcarrier is allocated to the user with the best channel gain. However, the computation complexity of these resource scheduling algorithms to find the optimal solution, such as maximizing sum rate or minimizing total transmission power, is tremendous. This complexity makes the algorithms impossible to be implemented in real time and renders them infeasible. Accordingly, low-complexity resource scheduling algorithms that find suboptimal solutions are investigated. Kulkarni *et al.* [5] studied a centralized power and bit loading algorithm for point-to-point OFDM networks and further proposed a distributed resource scheduling algorithm for wireless ad hoc networks. Kim *et al.* [6] proposed a heuristic algorithm that minimizes the user's received OFDM symbols to minimize the power consumption of the mobile station. The performance of the heuristic algorithm is near optimal. Mao and Wang [7] proposed a branch-and-bound (BnB)-based radio resource allocation (RRA) scheme for OFDMA systems and further designed a low-complexity BnB-based RRA scheme by preassignment and reassignment to obtain a suboptimal solution. Lee and Chong [8] studied an equal power allocation (EPA) scheme to reduce the complexity of resource scheduling in frequency-selective fading channels for OFDMA downlink systems. The authors showed that the EPA scheme attains near optimal performance when considering system throughput maximization.

To further reduce computational complexity, a kind of two-step resource allocation (RA) scheme was proposed. Peng *et al.* [9] studied a mixed tabu–greedy algorithm for broadcasting systems. In the first step, the algorithm applies a tabu search approach to search a feasible transmission time slot for each station. In the second step, a greedy method is used to maximize the throughput. Yang *et al.* [10] proposed a hybrid scheduling scheme for single carrier systems. This scheme combines a conventional multiuser selection scheme and a proportional fair scheduling scheme to provide flexible tradeoff between system capacity and fairness. Suh and Mo [11] investigated a two-step RA algorithm for multicast systems. The algorithm separates RA into subcarrier allocation and bit

Manuscript received March 30, 2011; revised August 8, 2011 and November 11, 2011; accepted January 17, 2012. Date of publication February 3, 2012; date of current version March 21, 2012. The review of this paper was coordinated by Prof. R. Jäntti.

The authors are with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: cjchang@mail.nctu.edu.tw).

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

Digital Object Identifier 10.1109/TVT.2012.2186832

loading to reduce the high computation complexity. However, these RA schemes do not consider quality-of-service (QoS) requirement guarantee. Che *et al.* [12] proposed a two-step power and channel allocation scheme for cognitive networks, which considered minimum rate constraints and fairness. Dai *et al.* [13] investigated a power allocation scheme for code-division multiple-access (CDMA) systems with minimum rate constraints. By splitting the max–min problem into two successive steps, the computational complexity of the power allocation was significantly reduced.

To support multimedia service, optimal RRA problems with QoS requirement constraints were studied. Zhang and Leung [14] proposed a resource scheduling scheme designed for nonreal-time (NRT) services to maintain predefined target rates. The proposed algorithm provides statistically proportional rates for NRT users and improves system throughput. With the delay constraint for real-time (RT) service, Wang and Chen [15] investigated an optimal resource partition algorithm that divides time slots and subchannels into two portions: one for random access and the other for connection-oriented access. Using a mixed-integer nonlinear programming technique, the proposed algorithm can determine the optimal amount of reserved time slots and subchannels for random access and maximize the overall efficiency of radio resource.

Given the increasing demands for multimedia mobile Internet services, researchers have also investigated RRA schemes that support differentiated QoS requirements for various service types, such as packet delay, packet dropping rate, and average transmission rate. An adaptive RRA (ARRA) scheme designed for RT, NRT, and best-effort (BE) mixed services was proposed in [16]. The ARRA scheme maximizes the system throughput under QoS requirement constraints with low computation complexity. The ARRA scheme performs much better than the multiuser maximum sum rate scheme in [17] and the truncated generalized processor sharing scheme in [18] on both system throughput and QoS guarantees. Wang and Dittmann [19] studied a two-level hierarchical RA scheme consisting of an aggregate resource allocator and multiple class schedulers for differentiated service classes. The aggregate resource allocator adjusts the amount of reserved resources for each service class, whereas each class scheduler allocates the reserved resources to users based on users' priority. Katoozian *et al.* [20] proposed a utility-based RRA (URRA) scheme for RT and BE services. The utility function of the URRA scheme takes packet delay and link quality into account to improve the QoS guarantee and system throughput. A utility-based throughput maximization and complexity reduction scheduling (U-TMCR) scheme for OFDMA downlink systems was also proposed in [21]. The U-TMCR scheme employs a heuristic allocation algorithm to reduce complexity and maximizes overall utility. This approach achieves high system throughput and low computational complexity.

These conventional RRA schemes previously mentioned are considered to balance the tradeoff between QoS requirement guarantee and system throughput enhancement. However, they mixed the two functions for QoS guarantee and throughput enhancement together; the way to balance this tradeoff is not explicitly expressed, and the tradeoff cannot be well controlled. How to design a scheduling scheme that can explicitly handle the balance and then achieve an excellent balance would be an essential issue. Note that achieving excellent balance helps maximize the system throughput under the QoS requirement fulfillment.

This paper proposes a new balanced resource scheduling (BRS) scheme with adaptive priority threshold to strike an excellent balance between QoS guarantee and system throughput for downlink transmission in OFDMA systems. The BRS scheme with adaptive priority thresholds is a two-stage resource scheduling scheme. It first performs a priority-based RA algorithm for users whose priority value is larger than their corresponding priority threshold to guarantee QoS requirement in the first stage. In the second stage, it performs a channel state information (CSI)-based RA algorithm for the remaining users to enhance system throughput. The BRS scheme also determines a priority threshold for each user to decide by which algorithm the user will be served. If a user's priority value is larger than its priority threshold, the user will be served by the priority-based RA algorithm in the first stage and by the CSI-based RA algorithm in the second stage otherwise. Moreover, a fuzzy inference system intelligently determines the priority threshold. A fuzzy inference priority threshold generator (FIPG) adaptively adjusts the priority threshold of each user by considering the QoS fulfillment with respect to QoS requirements and the channel status with respect to system throughput. Simulation results show that the BRS scheme with adaptive priority thresholds can enhance system throughput and guarantee QoS requirements. The BRS scheme achieves system throughput higher than the ARRA [16], URRA [20], U-TMCR [21], and fairness and QoS guarantee scheduling with fuzzy controls (FQFC) [27] schemes by 16%, 8.5%, 8.2%, and 46.8% at traffic load of 0.93, respectively, while still fulfilling the QoS requirements. The BRS scheme with adaptive priority thresholds adjusted by the FIPG also outperforms the BRS scheme with fixed priority thresholds in both QoS guarantee and throughput enhancement.

This paper is organized as follows. Section II describes the system model. Section III introduces the proposed BRS scheme. Section IV describes the FIPG. Sections V and VI present the simulation results and conclusions, respectively.

II. System Model

A. System Configuration

Fig. 1 depicts an OFDMA downlink system containing a two-tier 19-cell environment. Here, a frequency reuse pattern 3 is assumed so that the adjacent cells work on different frequency bands to mitigate intercell interference. In an OFDM frame, the frequency band and the time axis are divided into N subchannels (n_s subcarrier per subchannel) and L OFDM symbols, respectively. There are K active mobile users uniformly distributed in a cell. Base stations (BSs) and users are equipped with a single antenna. The system adopts an EPA [8] and supports adaptive modulation orders of M quadratic-amplitude modulation (QAM), where $M \in \{4, 16, 64\}$. The BRS scheme is performed frame by frame at the BS, and a basic

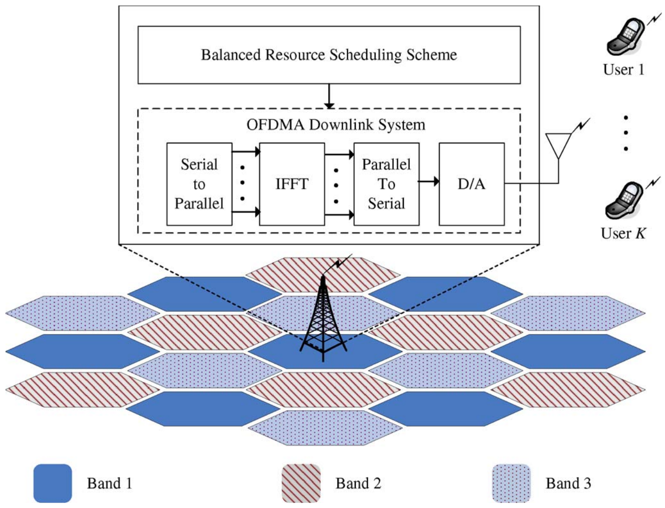

Fig. 1. Multicell environment for OFDMA downlink system.

resource block (RB) contains one subchannel and one OFDM symbol.

B. Multimedia Service Traffic

The OFDMA system can carry various classes of multimedia service traffic: 1) RT service; 2) NRT service; and 3) BE service. Voice and video traffic are both RT services, whereas hypertext transport protocol (HTTP) and file transfer protocol (FTP) traffic are NRT and BE services, respectively. This paper assumes that each active user has only one downlink service traffic. The maximum required bit error rate (BER) of user k, denoted by BER_k^*, is based on the traffic type belonging to user k as well. The QoS requirements for user k with RT service are the maximum delay tolerance, which is denoted by D_k^*, and the maximum allowable packet dropping rate, which is denoted by V_k^*. A packet of RT users will be dropped if its delay exceeds the specified maximum delay tolerance. For user k with NRT service, the QoS requirement is the minimum required transmission rate denoted by R_k^*. There is no specified QoS requirement for BE users.

C. CSI

This paper assumes that the coherent time of wireless channel is larger than the frame duration, and that the channel is fixed in a frame duration. Because the EPA for all subchannels is adopted, the transmission power on each subchannel is $P_{\max}/N$, where $P_{\max}$ is the maximum transmission power constraint of the BS. Assume that BS 0 is the anchor BS of user k, and denote $I_k^{(n)}$ as the interference power to user k on subchannel n from the other 18 cells in the first and second tiers. The value of $I_k^{(n)}$ can be obtained by

$$I_k^{(n)} = \sum_{i=1}^{18} \sum_{j \in \mathcal{K}_i} H_{i,k}^{(n)} \sqrt{p_{i,j}^{(n)}}\, s_{i,j}^{(n)} \tag{1}$$

where $H_{i,k}^{(n)}$, $p_{i,k}^{(n)}$, and $s_{i,k}^{(n)}$ denote the frequency-domain channel gain, the pilot power, and the pilot data symbol from BS i to user k on subchannel n, respectively. The term $\mathcal{K}_i$ denotes the set of users residing in cell i (BS i). The $H_{i,k}^{(n)}$ is given by

$$H_{i,k}^{(n)} = 10^{\{(\eta_{i,k}+\rho_{i,k})/10\}} \sum_{l=1}^{L_t} h_{k,i,l} e^{-j2\pi f_n \tau_l} \tag{2}$$

where $\eta_{i,k}$ is the path loss from BS i to user k, L_t is the number of taps of multipath channel, $\rho_{i,k}$ is the shadowing effect between BS i and user k, which is lognormal distributed with standard deviation σ_ρ, $h_{i,k,l}$ denotes the Rayleigh-faded attenuation of the lth tap between BS i and user k in time domain, f_n is the frequency of subchannel n, and τ_l is the delay spread of the lth tap. The received signal of user k on subchannel n at the current frame, which is denoted by $y_k^{(n)}$, is given by

$$y_k^{(n)} = H_{0,k}^{(n)} \sqrt{p_{0,k}^{(n)}}\, s_{0,k}^{(n)} + Z_k^{(n)} + I_k^{(n)} \tag{3}$$

where $Z_k^{(n)}$ is the thermal noise, which is assumed to be complex Guassian with zero mean and variance σ^2. Therefore, the signal-to-interference-plus-noise ratio (SINR) on subchannel n received at user k, denoted by $\text{SINR}_k^{(n)}$, can be obtained by

$$\text{SINR}_k^{(n)} = \frac{p_{0,k}^{(n)} |H_{0,k}^{(n)}|^2}{\sigma^2 + |I_k^{(n)}|^2}. \tag{4}$$

The minimum required SINR by applying M-QAM modulation order while satisfying BER_k^*, which is denoted by $\text{SINR}_k^*(M)$, is given by [22]

$$\text{SINR}_k^*(M) = \frac{\ln(5\text{BER}_k^*)(M-1)}{-1.5}, \qquad M \in \{4, 16, 64\}. \tag{5}$$

In this OFDMA downlink system, $\text{SINR}_k^{(n)}$ is regarded as the CSI of user k on subchannel n and will be quantized and reported by user k to the anchor BS via uplink channel every frame. Denote $\gamma_k^{(n)}$ as the quantized CSI of subchannel n reported by user k for the current frame. The term $\gamma_k^{(n)}$ then is defined as

$$\gamma_k^{(n)} = \begin{cases} 1, & \text{if } \text{SINR}_k^*(4) \le \text{SINR}_k^{(n)} < \text{SINR}_k^*(16) \\ 2, & \text{if } \text{SINR}_k^*(16) \le \text{SINR}_k^{(n)} < \text{SINR}_k^*(64) \\ 3, & \text{if } \text{SINR}_k^*(64) \le \text{SINR}_k^{(n)} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\gamma_k^{(n)} = 0$, 1, 2, and 3 indicate that the maximum supportable modulation orders are no transmission, quadrature phase-shift keying (QPSK), 16-QAM, and 64-QAM, respectively, on subchannel n of user k.

III. Balanced Resource Schuduling Scheme

The proposed BRS scheme strikes a balance between system throughput enhancement and QoS requirement guarantee and operates frame by frame. As shown in Fig. 2, the BRS scheme mainly consists of a priority-based RA algorithm and a CSI-based RA algorithm. At the beginning of every frame, it

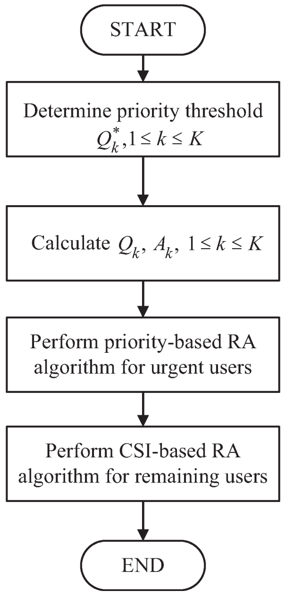
START
Determine priority threshold $Q_k^*, 1 \le k \le K$
Calculate Q_k, A_k, $1 \le k \le K$
Perform priority-based RA algorithm for urgent users
Perform CSI-based RA algorithm for remaining users
END

Fig. 2. Flowchart of the BRS scheme.

adaptively determines the priority threshold of user k, which is denoted by Q_k^*, and then calculates a priority value and an assigned rate of user k, which is denoted by Q_k and A_k, respectively, $1 \le k \le K$. According to Q_k^* and Q_k, user k is labeled as urgent user if $Q_k \ge Q_k^*$. The BRS scheme serves urgent users by the priority-based RA algorithm in the first stage to guarantee their QoS requirements. It then serves the remaining users by the CSI-based RA algorithm in the second stage to enhance system throughput. More importantly, Q_k^* is adaptively and intelligently adjusted by a fuzzy inference system. Such a configuration of the BRS scheme would be more flexible and better than those of the existing algorithms to well control the balance between system throughput enhancement and QoS requirement guarantee.

A. *Priority Value and Assigned Rate*

The priority value of user k, Q_k, is to denote how urgently user k requires for system resources to fulfill its QoS requirements. The term Q_k is defined as a basic priority value multiplied by an urgency term that exponentially increases (decreases) as the performance measure of user k with RT (NRT) service approaches its QoS requirement of maximum tolerable delay (minimum transmission rate). Assume that BE users have no strict QoS requirement and that the urgency term is set to be 1. Therefore, the Q_k is given by

$$Q_k = \begin{cases} q_{\text{RT}} \times \exp\left(\frac{D_k+1}{D_k^*}\right), & \text{if } k \text{ is a RT user} \\ q_{\text{NRT}} \times \exp\left(\frac{1.5R_k^* - 0.5\overline{R}_k}{R_k^*}\right), & \text{if } k \text{ is a NRT user} \\ q_{\text{BE}}, & \text{if } k \text{ is a BE user} \end{cases} \tag{7}$$

where q_{RT}, q_{NRT}, and q_{BE} are the basic priorities for RT, NRT, and BE services, respectively, and are set as $q_{\text{RT}} > q_{\text{NRT}} > q_{\text{BE}}$ to reflect the inherent priority differentiation among services. The term D_k is the QoS measure of delay that the head-of-line (HoL) packet of RT user k experiences, whereas D_k^* is the QoS requirement of the maximum tolerable delay for RT user k. The term $\overline{R}_k$ is the QoS measure of the average transmission rate of user k, whereas R_k^* is the QoS requirement of the minimum transmission rate for NRT user k. The constants 1.5 and 0.5 for Q_k of the NRT users in (7) are two shape factors that are used to tune the adaption degree of Q_k of NRT users. The factor 1.5 controls the largest value that Q_k can be, whereas the factor 0.5 controls the decreasing rate of Q_k when $\overline{R}_k$ increases. These shape factors are properly set according to the design expectation. Here, we expect that these factors make Q_k of NRT users at $\overline{R}_k = 0$ be larger than Q_k of RT users at $D_k \le D_k^*/2$ but be smaller than Q_k of RT users at $D_k = D_k$. They also make Q_k of NRT users at $\overline{R}_k = 2R_k^*/3$ be larger than Q_k of RT users at $D_k = 0$ but be smaller than Q_k of RT users at $D_k \ge D_k^*/2$. For RT users, to emphasize the urgency, Q_k increases rapidly when D_k is close to D_k^*. For NRT users, Q_k decreases when $\overline{R}_k$ is close to R_k^*, and it declines quickly when $\overline{R}_k$ exceeds R_k^*.

The assigned rate to user k, A_k, is the number of bits required by user k to keep the QoS requirement guaranteed. Under the just fulfillment of QoS requirement, A_k is gradually adjusted frame by frame according to the QoS measure of users. More specifically, if D_k of RT user k is lower than half of D_k^* (user k is not urgent), A_k is a portion, for example, D_k/D_k^*, of the residual HoL packet length of user k, which is denoted by B_k^H, and is the whole B_k^H otherwise. Likewise, if $\overline{R}_k$ of NRT user k is lower than R_k^* (user k violates the minimum rate constraint), A_k is given with $\min\{(R_k^* - \overline{R}_k)W_k, B_k\}$ bits to compensate for the low transmission rate, where W_k and B_k are the number of active frames and the remaining queuing length, respectively, of user k. Otherwise, A_k is given with a reasonable number of bits, for example, $\min(0.5R_k^*, B_k)$. Therefore, A_k can be formulated as

$$A_k = \begin{cases} \left\lceil \frac{D_k}{D_k^*} \times \frac{B_k}{r} \right\rceil \times r, & \text{if } k \in \text{RT}, D_k \le \frac{D_k^*}{2} \\ \left\lceil \frac{B_k}{r} \right\rceil \times r, & \text{if } k \in \text{RT}, D_k > \frac{D_k^*}{2} \\ \left\lceil \frac{\min\{(R_k^* - \overline{R}_k)W_k, B_k\}}{r} \right\rceil \times r, & \text{if } k \in \text{NRT}, \overline{R}_k < R_k^* \\ \left\lceil \frac{\min(0.5R_k^*, B_k)}{r} \right\rceil \times r, & \text{if } k \in \text{NRT}, \overline{R}_k \ge R_k^* \\ \left\lceil \frac{B_k}{r} \right\rceil \times r, & \text{if } k \in \text{BE} \end{cases} \tag{8}$$

where r is the minimum unit for RA, $r = 2n_s$ bits for QPSK modulation, and the term $\lceil x \rceil$ is the smallest integer larger than x.

B. *Priority- and CSI-Based RA Algorithms*

Given Q_k^*, Q_k, and A_k, $1 \le k \le K$, the BRS scheme first performs the priority-based RA algorithm for urgent users k whose priority value Q_k is higher than or equal to the priority threshold Q_k^*. When $Q_k \ge Q_k^*$, the urgent user k is in a risk of QoS requirement violation and should be served immediately. With the priority-based RA algorithm, urgent users k are served with a data amount of A_k in the service order of priority value Q_k. The best subchannel n with the highest $\gamma_k^{(n)}$ is assigned to the selected urgent user k. The priority-based RA algorithm stops when the allocated rates, which is denoted by R_k, of urgent users k are all satisfied with A_k. The BRS scheme then

TABLE I
DEFINITION OF VARIABLES

D_k^*	Maximum tolerable packet delay of user k
D_k	HoL packet delay of user k
V_k^*	Maximum allowable packet dropping rate of user k
V_k	Packet dropping rate of user k
R_k^*	Minimum required trans. rate of user k
$\overline{R}_k$	Average trans. rate of user k
Q_k^*	Priority threshold of user k
Q_k	Priority value of user k
ΔQ_k^*	Adjustment step of the priority threshold of user k
r	Minimum unit for resource allocation
$\gamma_k^{(n)}$	Quantized SINR (CSI) of user k on subchannel n
$V_k^{(-)}$	Packet dropping rate of user k in the last frame
$R_k^{(-)}$	Average trans. rate of user k in the last frame
$Q_k^{*(-)}$	Priority threshold of user k in the last frame

Note that all variables are defined for the current frame unless they are specified to the last frame.

performs the CSI-based RA algorithm for the remaining users to enhance the system throughput. With the CSI-based RA algorithm, user k with the highest $\gamma_k^{(n)}$ on subchannel n is first scheduled and allocated an amount of remaining HoL packet length, denoted by B_k, for throughput enhancement. If more than one user k on subhchannel n exhibits the same $\gamma_k^{(n)}$, the one with the highest Q_k is allocated first. Both RA algorithms stop when resources are exhausted or no user has data in the queue. The pseudocode in the Appendix describes the details of both RA algorithms.

Notice that in the BRS scheme with adaptive priority thresholds, user k with the higher priority threshold is more likely to be served by the CSI-based RA algorithm, whereas user k with the lower priority threshold has a higher chance to be served by the priority-based RA algorithm. Thus, the BRS scheme can well control the tradeoff between throughput enhancement and QoS guarantee using adaptively determining the priority threshold of each user. Finally, we provide a list of variables that will be used throughout this paper in Table I.

IV. FUZZY INFERENCE PRIORITY THRESHOLD GENERATOR

To strike an excellent balance between system throughput enhancement and QoS requirement guarantee, the priority threshold of each user, which is an essential parameter in the BRS scheme, must be intelligently and adaptively determined. However, the relationship between system throughput and QoS performance in the OFDMA downlink system is complicated, uncertain, and difficult to model mathematically. Therefore, it is impossible by hard computation methods to get an appropriate priority threshold to achieve an excellent balance. Fortunately, fuzzy inference systems can efficiently solve such uncertain, time varying, nonlinear problems [23]. A fuzzy inference system is an improved and intelligent design that utilizes the mathematical formulation of classical control to mimic the expert knowledge [23]. A fuzzy inference system also provides effective solutions with linear computation complexity related to the number of fuzzy inference rules. Other effective intelligent techniques include neural fuzzy system and fuzzy Q-learning [23]. However, because of the high efficiency, low complexity, and easy implementation of a fuzzy inference system, this paper adopts a fuzzy inference system to determine the priority threshold of each user.

For decades, fuzzy inference systems have been applied to wireless communication systems to adapt to the complicated and uncertain environments [24]–[27]. A fuzzy channel allocation controller for hierarchical cellular systems was proposed in [24]. Ye *et al.* designed a call admission control (CAC) scheme for uplink transmission in wideband CDMA systems based on fuzzy logics [25]. Tsay *et al.* proposed a fuzzy power control scheme for downlink transmission in CDMA-based local multipoint distribution service systems [26]. Chen *et al.* designed a fairness and QoS guarantee scheduling scheme with fuzzy control [27]. All can achieve better performance.

This paper designs a sophisticated FIPG to intelligently determine the priority threshold of each user. The FIPG consists of a fuzzifier, a fuzzy rule base, an inference engine, and a defuzzifier. Three input linguistic variables of user k are effectively selected: M_k, S_k, and C_k. The output linguistic variable is the adjustment step of Q_k^*, which is denoted by ΔQ_k^*. To adapt to different user states, the size of each adjustment step is variable just like the adaptive delta modulation [28], depending on the values of M_k, S_k, and C_k.

The term M_k denotes the QoS status indicator of user k with respect to the QoS requirement at the current frame and is denoted as

$$M_k = \begin{cases} \frac{V_k^* - V_k}{V_k^*}, & \text{if } k \text{ is a RT user} \\ \frac{\overline{R}_k - R_k^*}{R_k^*}, & \text{if } k \text{ is a NRT user} \end{cases} \tag{9}$$

where V_k is the QoS measure of the packet dropping rate of user k with voice or video services in the current frame, and V_k^* is the QoS requirement of the packet dropping rate. If M_k is positive (negative), the QoS requirement of user k is guaranteed (violated), and ΔQ_k^* should be positive (negative) to increase (decrease) Q_k^* for the sake of throughput enhancement (QoS guarantee).

The S_k denotes the QoS tendency indicator of user k in the current frame with respect to the last frame and can be written as

$$S_k = \begin{cases} \frac{V_k^{(-)} - V_k}{\max\left(V_k^{(-)}, \epsilon_V\right)}, & \text{if } k \text{ is a RT user} \\ \frac{\overline{R}_k - \overline{R}_k^{(-)}}{\max\left(\overline{R}_k^{(-)}, 1\right)}, & \text{if } k \text{ is a NRT user} \end{cases} \tag{10}$$

where $V_k^{(-)}$ and $\overline{R}_k^{(-)}$ are the packet dropping ratio and the average transmission rate of user k measured at the last frame, respectively, and ϵ_V is a tiny number less than 10^{-6} to avoid the case of dividing by zero. Unlike M_k, the term S_k indicates the variation tendency of QoS measures. If $S_k \geq 0$, it means that the QoS measure of the packet dropping ratio (average transmission rate) of user k is improves or remains unchanged, and the QoS measure of user k tends to be stable. Otherwise, the QoS measure of user k should be monitored because it tends to deteriorate.

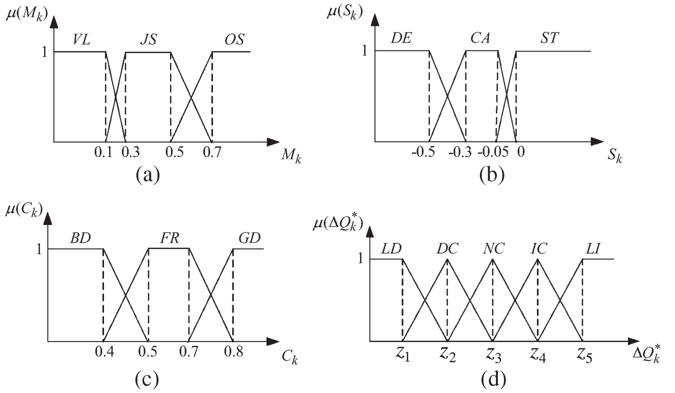

Fig. 3. Membership functions of the fuzzy linguistic variables. (a) Membership functions of M_k. (b) Membership functions of S_k. (c) Membership functions of C_k. (d) Membership functions of ΔQ_k^*.

The C_k denotes the overall channel condition indicator of user k at the current frame and is given by

$$C_k = \frac{\sum_{n=1}^{N} \gamma_k^{(n)}}{3N} \tag{11}$$

where the denominator $3N$ stands for that all N subchannels can support the highest modulation order 64-QAM ($\gamma_k^{(n)} = 3$). If C_k is high (low), the majority of the subchannels of user k is in a good (bad) channel condition, and user k can be assigned these subchannels with a high modulation order.

Term sets for the input variables M_k, S_k, and C_k in the fuzzifier are designated as $T(M_k) = \{OS(Over\text{-}Satisfied), JS(Just\text{-}Satisfied), VL(Violated)\}$, $T(S_k) = \{ST(Stable), CA(Cautionary), DE(Deteriorative)\}$, and $T(C_k) = \{GD(Good), FR(Fair), BD(Bad)\}$. Membership functions of each linguistic term for M_k, S_k, and C_k are trapezoidal, as Fig. 3(a)–(c), respectively, show. To set the edges of the trapezoidal functions, consider the example of M_k in Fig. 3(a). Set $(-\infty, 0.1)$ ($(0.7, \infty)$) for term $VL(OS)$ since the QoS status of RT user k, M_k, is regarded as violated (oversatisfied) if the packet dropping rate of user k exceeds 90% (less than 30%) of the required dropping rate.

The linguistic term set of ΔQ_k^* is defined as $T(\Delta Q_k^*) = \{LD(Largely\ Decreased), DC(Decreased), NC(No\ Changed), IC(Increased), LI(Largely\ Increased)\}$, where the terms in $T(\Delta Q_k^*)$ describe the adjustment step size. Fig. 3(d) depicts the membership function of each linguistic term. Set $z_3 = 0$ to indicate the step size of no change. Set z_m to $(m-3) \times \nu$, $1 \le m \le 5$ so that z_1 and z_2 (z_4 and z_5) represent various negative (positive) step sizes, where ν is the minimum step size for the adjustment. The ν is defined as

$$\nu = \frac{Q_{\max} - Q_{\min}}{X} \tag{12}$$

where $Q_{\max}$ ($Q_{\min}$) is the upper (lower) limit of the priority threshold, and the constant X is the step size resolution.

Table II lists the fuzzy rules of the FIPG, where the rules are based on the domain knowledge and explained as follows. As mentioned earlier, increasing the priority threshold of a user increases the chances of the user being served by the CSI-based RA algorithm, which enhances the system throughput. Conversely, decreasing the priority threshold of a user increases the possibility of that user being served by the priority-based RA algorithm, which guarantees QoS requirement. Thus, if the QoS measure of user k is oversatisfied, user k can be served by the CSI-based RA algorithm to enhance the system throughput without a QoS violation. Therefore, in rules 1–3, ΔQ_k^* is (largely) increased, regardless of the QoS tendency indicator S_k. On the other hand, if the QoS measure is violated, in rules 12 and 13, ΔQ_k^* decreases regardless of the overall channel condition indicator C_k. This increases the opportunity of being scheduled by the priority-based RA algorithm. However, if the QoS measure is violated but tends toward stable due to the previous adjustment, ΔQ_k^* is designed to remain unchanged to retain the current QoS tendency and avoid continuous priority threshold declining, as given in rule 11.

TABLE II
FUZZY RULE BASE OF FIPG

Rule	M_k, S_k, C_k	ΔQ_k^*	Rule	M_k, S_k, C_k	ΔQ_k^*
1	*OS, −, GD*	*IC*	8	*JS, CA, FR*	*NC*
2	*OS, −, FR*	*LI*	9	*JS, CA, BD*	*DC*
3	*OS, −, BD*	*LI*	10	*JS, DE, −*	*DC*
4	*JS, ST, GD*	*IC*	11	*VL, ST, −*	*NC*
5	*JS, ST, FR*	*IC*	12	*VL, CA, −*	*DC*
6	*JS, ST, BD*	*LI*	13	*VL, DE, −*	*LD*
7	*JS, CA, GD*	*NC*			

− : Do Not Care.

TABLE III
SYSTEM-LEVEL PARAMETERS

Parameters	Value
Cell size	1.6 km
Frame duration	5 ms
System bandwidth	5 MHz
FFT size	512
Subcarrier frequency spacing	10.9375 KHz
No. of data subcarriers	384
No. of subchannels (N)	8
No. of data subcarriers per subchannel (n_s)	48
No. of slots for downlink per frame (L)	20
Maximum transmission power (P_{max})	43 dBm
Thermal noise density	-174 dBm/Hz

The inference engine adopts the max–min inference method, and the defuzzifier uses the center of area defuzzification method [23] to obtain the crisp value of ΔQ_k^*. After ΔQ_k^* is obtained, Q_k^* for the current frame is updated by

$$Q_k^* = \begin{cases} \min(Q_k^{*(-)} + \Delta Q_k^*, Q_{\max}), & \text{if } \Delta Q_k^* \ge 0 \\ \max(Q_k^{*(-)} + \Delta Q_k^*, Q_{\min}), & \text{if } \Delta Q_k^* < 0 \end{cases} \tag{13}$$

where $Q_k^{*(-)}$ is the priority threshold of user k determined in the last frame.

V. Simulation Results

A. Simulation Environment

Table III lists the system-level parameters, where the parameters of the physical layer are configured according to the suggested values in [29]. Large- and small-scale fading are considered in the wireless fading channel, where the large-scale

TABLE IV
QoS REQUIREMENTS OF EACH TRAFFIC TYPE

QoS Requirements	voice (RT)	video (RT)	HTTP (NRT)	FTP (BE)
Required BER (BER$_k^*$)	10^{-3}	10^{-4}	10^{-6}	10^{-6}
Max. Packet Delay Tolerance (D_k^*)	40 ms	20 ms	N/A	N/A
Max. Packet Dropping Ratio (V_k^*)	1%	1%	N/A	N/A
Min. Required Trans. Rate (R_k^*)	N/A	N/A	100 kbps	N/A

N/A : Not Applicable.

fading is caused by path loss and shadowing effect, and the small-scale fading comes from multipath reflection. The path loss from BS i to user k in (2), i.e., $\eta_{i,k}$, is modeled as $128.1 + 37.6 \log d_{i,k}$ [30], where $d_{i,k}$ is the distance between the BS i and user k measured in kilometers. The standard deviation σ_ρ of the lognormal shadowing $\rho_{i,k}$ for (2) is assumed to be 8 dB. The Rayleigh-faded multipath channel $h_{i,k,l}$ is modeled with six Rayleigh-faded taps $(L_t = 6)$, and the power delay profile is defined as {1, 0.60653, 0.36788, 0.22313, 0.13534, 0.082085} following the exponential decay rule. The channel model for each user is independent and identically distributed, and the channel is assumed to remain fixed within a frame while varying from frame to frame. The mobility of users is randomly distributed between 0 and 60 km. The simulation platform is developed by C++.

B. Traffic Model and QoS Requirement

The traffic models of voice, video, HTTP, and FTP are described as follows. The voice traffic is modeled as an ON–OFF model, in which the lengths of the ON and OFF periods follow an exponential distribution with means 1.0 and 1.35 s [31], respectively. During the ON period, a packet is generated every 20 ms. With a 8-kb/s voice encoder rate, the voice packet size is 28 B including payload and header. No packet is generated during the OFF period. The parameters of voice traffic are configured according to [32]. Streaming video packets are assumed to arrive at a regular interval of 100 ms. Each video frame is decomposed into eight slices (packets), and the size of the packet is truncated Pareto distributed with a mean of 100 B. The interval between packets is truncated Pareto distributed with a mean of 6 ms. For the HTTP traffic of NRT service, the behavior of web browsing is modeled as a sequence of page downloads, and each page download is modeled as a sequence of packet arrivals. Each page consists of a main object and several embedded objects whose packet size is truncated lognormal distributed with a mean of 10 710 and 7758 B, respectively. Both main and embedded objects are divided into several packets with a maximum transmission unit of 1500 B. The interval between two consecutive page downloads, representing the reading time in web browsing, is distributed in an exponential distribution with a mean of 30 s. The FTP traffic of the BE service is modeled as a sequence of file downloads. The size of a file is distributed in a truncated lognormal distribution with a mean of 2M B. In addition, the interval between files is distributed in an exponential distribution with a mean of 180 s. The parameters of video streaming, HTTP, and FTP traffic are configured according to [30]. Table IV lists the QoS requirements of each traffic type [33]–[36].

The simulations in this paper assume that the number of users in the four traffic types is the same. The traffic intensity of the system is defined as the total average arrival rate of all traffic over the maximum system transmission rate, where the maximum system transmission rate is achieved when all RB s are allocated with the highest modulation order. Since the average date rates of voice, video, HTTP, and FTP traffic id are 4.8, 64, 14.5, and 88.9 kb/s, respectively, the traffic intensity varies from 0.18 to 0.93 as the number of users varies from 40 to 200.

C. Performance Evaluation

This section compares the proposed BRS scheme with adaptive priority threshold with four conventional schemes: the ARRA scheme in [16], the URRA scheme in [20], the U-TMCR scheme in [21], and the FQFC scheme in [27]. The parameters of the BRS scheme are configured as follows: $q_{\mathrm{RT}} = 3.5$, $q_{\mathrm{NRT}} = 2$, $q_{\mathrm{BE}} = 1$, $Q_{\max} = 12$, $Q_{\min} = 2$, and $X = 10$. The ARRA scheme is a kind of a two-step RA scheme designed for OFDMA downlink systems with multimedia traffic. The ARRA scheme is composed of two parts to solve the optimization problem of RRA. The first part is a dynamic priority adjustment algorithm, where the priorities of users are dynamically adjusted, based on the time-to-expiration value and the remaining packet size of the HoL packet, frame by frame. The second part is a priority-based greedy (PBG) algorithm. The intention of the PBG algorithm is to maximize the total system throughput under four system constraints. It uses the greedy principle to find the best allocation and can be considered as a joint design of power, subchannel, and bit allocation in the physical layer. The ARRA scheme determines the priority value for users based on the time-to-expire value and the remaining packet size of the HoL packet and serves users in the order of priority value. Notice that, in the ARRA scheme, NRT users may have priority values higher than RT users because NRT users usually have larger packets. The URRA and U-TMCR schemes schedule users by maximizing the overall utility value, where the utility value consists of a QoS factor and a data rate factor of users. However, the URRA scheme does not consider the NRT service and sets the value of the QoS factor in between 1 and 8 for RT users and 1 for BE users. To make a fair comparison, a QoS factor is defined as $(1 + R_k^*/R_k^* + \overline{R}_k)$ given to NRT users in the URRA scheme. The rationales of the definition of the QoS factor are as follows: 1) The NRT users with low average transmission rate will get a large QoS factor. 2) The QoS factor of NRT users does not exceed that of RT users when RT and NRT users are both urgent. 3) NRT users must have a larger QoS factor than BE users, and NRT users must have a higher serving opportunity than BE users. Note that these design principles of the QoS factor for NRT

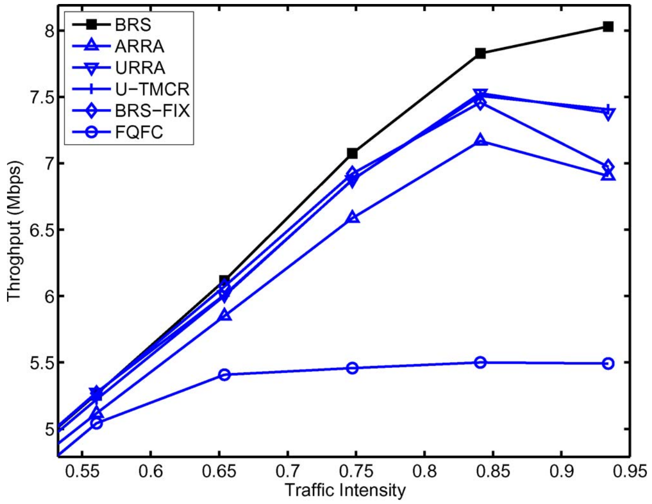

Fig. 4. System throughput.

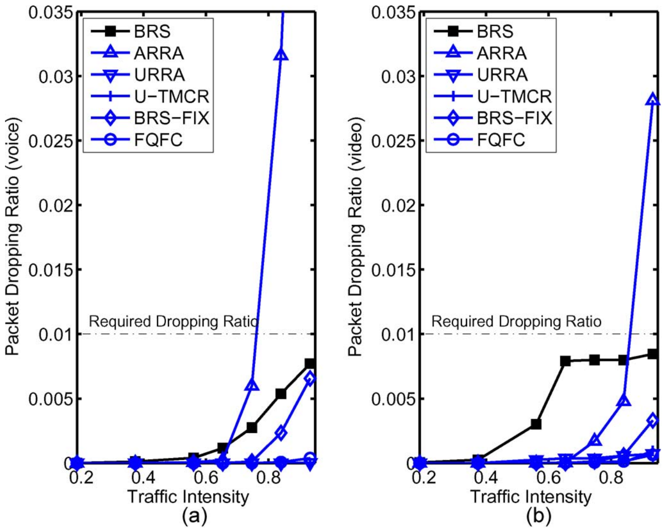

Fig. 5. Average packet dropping rate of RT users.

users of the URRA scheme are similar to the QoS factor of the U-TMCR and the priority value of the proposed BRS scheme. The FQFC scheme serves users in order of priority. It employs a fuzzy controller to adjust the priority values of RT users to have the HoL packet delay approach the goal delay of users. The goal delay is also adjusted by a fuzzy controller based on the system load and channel condition. To make the FQFC scheme operate properly in the considered system of this paper, a subchannel assignment mechanism, which assigns the selected serving users with their best subchannels(s), is assumed in the FQFC scheme. Besides the four conventional schemes, the BRS scheme with fixed priority thresholds, which is labeled as BRS-FIX, is also compared. The priority thresholds of the BRS-FIX are set to be 8 and 6 for RT users and NRT users, respectively.

Fig. 4 depicts the system throughput, indicating that the BRS scheme achieves the highest system throughput. The BRS scheme enhances system throughput 16%, 8.5%, 8.2%, 46.8%, and 14.9% at a traffic intensity of 0.93, compared with the ARRA, URRA, U-TMCR, FQFC, and BRS-FIX schemes, respectively. This is because the BRS scheme strikes a balance between the QoS requirement fulfillment (by the priority-based RA algorithm) and the system throughput enhancement (by the CSI-based RA algorithm) with adaptive priority thresholds that are intelligently determined by the sophisticated FIPG. As the design of FIPG given in Section IV, the FIPG effectively selects the QoS status indicator, the QoS tendency indicator, and the overall channel condition indicator as input linguistic variables. These input variables can sufficiently determine the adjustment step of the priority threshold. In addition, the fuzzy rules of the FIPG are designed based on throughput enhancement with proper QoS guarantee. For example, if the QoS measure is oversatisfied, the FIPG increases the priority threshold according to the overall channel condition indicator so that the user is more likely to be served by the CSI-based RA algorithm to enhance throughput. On the other hand, if the QoS measure is violated, the FIPG gradually decreases the priority threshold according to the QoS tendency indicator to prevent serious throughput degradation. Therefore, the BRS scheme strikes an excellent balance between QoS guarantee and throughput enhancement. In contrast, the throughput of the ARRA scheme decreases significantly as the traffic intensity increases. This is because the ARRA scheme allocates resources to the high priority users before the high SINR users. Hence, the spectrum efficiency of the ARRA scheme is low. Moreover, the URRA and U-TMCR schemes dedicate more resources than the BRS scheme to maintain perfect packet dropping rates for RT users, causing throughput degradation. For the BRS-FIX scheme, the throughput is also decreased at high traffic intensity. This is because the number of users served by the priority-based (CSI-based) RA algorithm increases (decreases) as traffic load becomes heavy. The FQFC scheme performs the worst and its system throughput is saturated at a traffic intensity of 0.65 because the FQFC scheme serves RT and NRT users in order of priority rather than channel condition. Moreover, it serves BE users in a round-robin fashion to attain fairness, which causes poor spectrum efficiency and seriously degrades the system throughput.

Fig. 5(a) and (b) shows the average packet dropping rates of voice and video users, respectively. The packet dropping rates of RT users by the BRS scheme are well controlled and tend toward just satisfied. The FIPG of the BRS scheme increases the priority threshold of users whose packet dropping rate is much lower than the required packet dropping rate. Thus, users are likely to be served by the CSI-based RA algorithm to enhance system throughput. However, if the packet dropping rate is close to the required dropping rate, the FIPG stops increasing or even decreases the priority threshold as necessary. Consequently, the BRS scheme prevents the packet dropping rates from oversatisfied and enhances the system throughput using adaptive priority thresholds. Conversely, the packet dropping rates of the ARRA scheme dramatically increase as the number of users increases. This is because, as previously mentioned, the priority of the ARRA scheme is partly related to the remaining HoL packet size of users, and RT users may get lower priority than NRT users. This is because NRT users usually have a packet size much larger than RT users. Moreover, the URRA, U-TMCR, and FQFC schemes have almost zero voice and video packet dropping rates, which

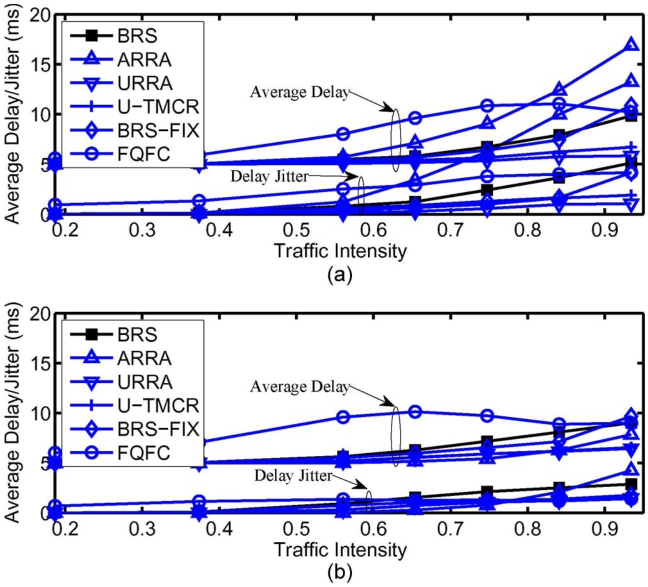

Fig. 6. Average delay and delay jitter of RT users.

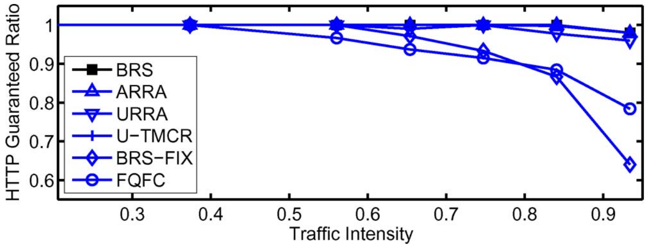

Fig. 7. Guaranteed ratio of NRT users.

shows that these schemes put too much emphasis on the QoS requirement guarantee for RT users.

Fig. 6 shows the average packet delay and delay jitter performance of voice/video users. The FQFC scheme has a higher average packet delay because it controls the HoL packet delay of RT users to approach the predetermined goal delay. It can be seen that the average delay and the jitter performance of voice/video users by the proposed BRS scheme are a little bit higher than those by some existing algorithms. It is because the design objective of the BRS scheme is to maximize the system throughput under the just fulfillment of the QoS requirement. The BRS scheme will put off serving the RT user with a poor channel gain until the RT user has a favorable channel condition or a priority value larger than the priority threshold. Therefore, the BRS scheme can achieve the highest system throughput among the compared schemes, as shown in Fig. 4, and attain the average packet dropping rate of RT users still under the QoS requirement of 0.01, as shown in Fig. 5.

Fig. 7 depicts the guaranteed ratio of NRT users, which is defined as the ratio of the number of NRT users whose average transmission rates are not less than the minimum transmission rate to the total number of NRT users. As the traffic load increases, the BRS, ARRA, URRA, and U-TMCR schemes can still maintain a guarantee ratio exceeding 95%. However, the HTTP guaranteed ratio of the FQFC scheme decreases when the traffic intensity increases and goes down to 78% at a traffic load of 0.93 because the FQFC scheme gives the unsatisfied NRT users with a fixed priority and assigns them a constant amount of resources. With the fixed priority, it cannot adequately reflect the urgency among the unsatisfied NRT users. With the constant amount of allocated resources, it could be insufficient for some unsatisfied NRT users to compensate their low transmission rates to maintain the minimum required transmission rate. The guaranteed ratio of the BRS-FIX scheme decreases to 64% when the traffic intensity is 0.93. This is because the priority thresholds for NRT users in the BRS-FIX scheme are fixed at 6; these priority thresholds cannot be adaptively adjusted, even when these NRT users are not satisfied with the minimum transmission rate.

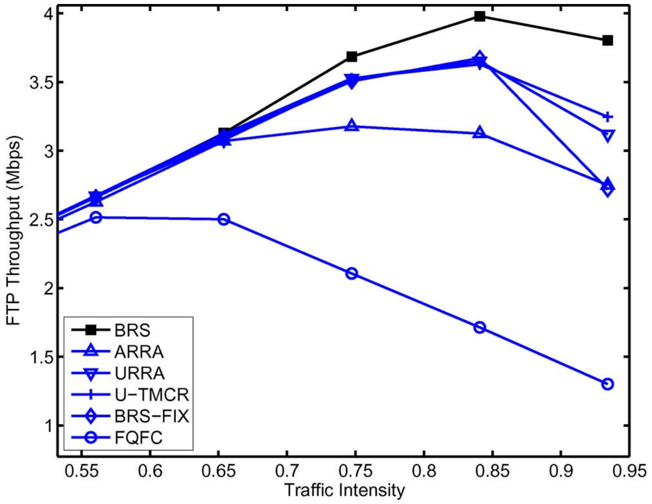

Fig. 8. Average throughput of BE users.

Fig. 8 shows the average throughput of BE users. The FTP throughput of all schemes except FQFC significantly decreases when the traffic intensity exceeds 0.85. This is because, at high traffic load, more resources must be allocated to urgent service users to maintain their specific QoS requirements. As a result, there are fewer resources for the FTP users. However, the proposed BRS scheme attains an average FTP throughput higher than the ARRA, URRA, U-TMCR, and FQFC schemes by 37.8%, 21.3%, 16.5%, and 192%, respectively. The BRS scheme can effectively allocate resources to users by intelligently determining their priority thresholds. Using the two-stage scheduling scheme and adaptive priority thresholds, the BRS scheme can minimize the amount of resources allocated to users who have a high priority but poor channel condition and have still not violated their QoS requirements. Furthermore, the BRS scheme usually serves the voice, video, and HTTP users as they have favorable channel condition. Therefore, these users can be satisfied with corresponding QoS requirements using fewer resources and making more resources available for the FTP users. In contrast, other schemes dedicate more resources to the voice and video users or even HTTP users (ARRA) maintaining QoS requirements by sacrificing the throughput of the FTP users. It can also be seen that the FTP throughput of the FQFC scheme decreases a lot when the traffic intensity exceeds

0.65. It is because the FQFC scheme serves the BE users in a round-robin fashion and thus has poor spectrum efficiency and a saturated system throughput at a traffic intensity of 0.65, as shown in Fig. 4. When the traffic load becomes higher, similar to other schemes, the FQFC scheme remains fewer resources from the saturated total system throughput for the BE users. Thus, the FTP throughput of the FQFC scheme reduces.

Consider the computational complexity of the proposed BRS scheme in allocating the system NL RBs, where N is the number of subchannels, and L is the number of symbols (time slots). The priority-based RA algorithm takes $O(K+N)$ to find the most urgent user and the best subchannel for the user; the computational complexity of the priority-based RA algorithm is $O((K+N)NL)$. The CSI-based RA algorithm takes $O(KN)$ to search the best user–subchannel pair; the complexity of the CSI-based RA algorithm is $O(KN^2L)$. In addition, the complexity of FIPG for K users is $O(KJ)$, where J is the number of fuzzy rules of the FIPG. Overall, the computational complexity of the proposed BRS scheme is about $O(KN^2L)$, which is less than a 5-ms frame time.

VI. Conclusion

This paper has proposed a new BRS scheme with adaptive priority thresholds for OFDMA downlink systems. Unlike conventional RRA schemes, the BRS scheme employs an adaptive priority threshold for each user to accurately control the tradeoff between system throughput enhancement and QoS requirement guarantee. The proposed BRS scheme is a two-stage scheduling scheme. This scheme performs a priority-based RA algorithm in the first stage and a CSI-based RA in the second stage. The former is for those urgent users whose priority values are larger than or equal to the priority threshold to fulfill the QoS requirement. The latter is for the remaining users who have a good channel condition to achieve high system throughput. A fuzzy inference system intelligently determines the priority threshold of each user, achieving an excellent control of this tradeoff. The proposed FIPG considers three essential parameters related to the priority threshold. Simulation results show that the proposed BRS scheme enhances the system throughput by 16%, 8.5%, 8.2%, and 46.8% at a traffic load of 0.93, compared with the conventional ARRA, URRA, U-TMCR, and FQFC schemes, respectively, under a QoS requirement guarantee. The BRS scheme with adaptive priority thresholds also outperforms the BRS scheme with fixed priority thresholds in both throughput enhancement and QoS guarantee. The proposed BRS scheme is therefore an effective scheme that provides a low computation complexity solution to the problem of finding a balance between QoS guarantee and system throughput enhancement for OFDMA downlink systems. Moreover, to prevent too much overloading the system and violating QoS requirements of multimedia traffic, CAC is generally needed to manage the total number of users in the system. By providing information such as the amount of residual resource and QoS measures of existing calls, the proposed BRS scheme can work well with the CAC schemes, such as the schemes in [37] and [38].

Appendix
Pseudocode of the Priority- and CSI-Based RA Algorithms

```
Algorithm 1: PRIORITY_BASED_RA
$\mathcal{K}_{PRIO} \longleftarrow \{k | Q_k^* < Q_k, k \in \mathcal{K}\}$;
while $\mathcal{K}_{PRIO} \neq \emptyset$ and $\mathcal{N} \neq \emptyset$ do
    $k^* \longleftarrow \arg \max_{k \in \mathcal{K}_{PRIO}} Q_k$;
    $n^* \longleftarrow \arg \max_{n \in \mathcal{N}} \gamma_{k^*}^{(n)}$;
    for $\ell \in \mathcal{L}^{(n^*)}$ do
        call ResourceAllocation($k^*, n^*, \ell$);
        if $A_{k^*} \leq R_{k^*}$ or user $k^*$ has no queuing data
        then
            $\mathcal{K}_{PRIO} \longleftarrow \mathcal{K}_{PRIO} - k^*$;
            break;
```

```
Algorithm 2: CSI_BASED_RA
while $\mathcal{K} \neq \emptyset$ and $\mathcal{N} \neq \emptyset$ do
    $(\gamma_{\max}, n^*) \longleftarrow \arg \max_{\gamma_k^{(n)}, n \in \mathcal{N}} \gamma_k^{(n)},\ k \in \mathcal{K}$;
    $\mathcal{K}_{\max} \longleftarrow \{k | \gamma_k^{(n*)} = \gamma_{\max},\ k \in \mathcal{K}\}$;
    $k^* \longleftarrow \arg \max_{k \in \mathcal{K}_{\max}} Q_k$;
    for $\ell \in \mathcal{L}^{(n*)}$ do
        call ResourceAllocation($k^*, n*, \ell$);
        if $B_{k^*} \leq R_{k^*}$ or user $k^*$ has no queuing data
        then
            break;
```

```
Procedure ResourceAllocation
  Data: $k, n, \ell$
$R_k \longleftarrow R_k + r \times \gamma_k^{(n)}$;
$B_k \longleftarrow B_k - r \times \gamma_k^{(n)}$;
$\mathcal{L}^{(n)} \longleftarrow \mathcal{L}^{(n)} - \ell$;
if $\mathcal{L}^{(n)} = \emptyset$ then
    $\mathcal{N} \longleftarrow \mathcal{N} - n$;
if user $k$ has no queuing data then
    $\mathcal{K} \longleftarrow \mathcal{K} - k$;
```

References

[1] *Local and Metropolitan Area Networks-Part 16: Air Interface for Broadband Wireless Access Systems*, IEEE Std. 802.16, 2009.

[2] "3GPP TR36.201: Evolved Universal Terrestrial Radio Access (E-UTRA); Long Term Evolution (LTE) Physical Layer; General Description," Eur. Telecommun. Std. Inst., Sophia-Antipolis, France, Third-Generation Partnership Project Tech. Rep., 2009.

[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. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[4] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, Feb. 2003.

[5] G. Kulkarni, A. Adlakha, and M. Srivastava, "Subcarrier allocation and bit loading algorithms for OFDMA-based wireless networks," *IEEE Trans. Mobile Comput.*, vol. 4, no. 6, pp. 652–662, Nov. 2005.

[6] J. Y. Kim, T. S. Kwon, and D. H. Cho, "Resource allocation scheme for minimizing power consumption in OFDM multicast systems," *IEEE Commun. Lett.*, vol. 11, no. 6, pp. 486–488, Jun. 2007.

[7] Z. Mao and X. Wang, "Efficient optimal and suboptimal radio resource allocation in OFDMA system," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 440–445, Feb. 2008.

[8] H.-W. Lee and S. Chong, "Downlink resource allocation in multi-carrier systems: Frequency-selective vs. equal power allocation," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3738–3747, Oct. 2008.

[9] Y. Peng, B. H. Soong, and L. Wang, "Broadcast scheduling in packet radio networks using mixed tabu-greedy algorithm," *Electron. Lett.*, vol. 40, no. 6, pp. 375–376, Mar. 2004.

[10] L. Yang, M. Kang, and M.-S. Alouini, "On the capacity-fairness tradeoff in multiuser diversity systems," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1901–1907, Jul. 2007.

[11] C. Suh and J. Mo, "Resource allocation for multicast services in multicarrier wireless communications," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 27–31, Jan. 2008.

[12] Y. Che, J. Chen, W. Tang, and S. Li, "A two-step channel and power allocation scheme in centralized cognitive networks based on fairness," in *IEEE VTC-Spring*, May 2008, pp. 1589–1593.

[13] J. Dai, Z. Ye, and X. Xu, "Power allocation for maximizing the minimum rate with QoS constraint," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4989–4996, Nov. 2009.

[14] Y. Zhang and C. Leung, "Resource allocation for non-real-time services in OFDM-based cognitive radio systems," *IEEE Commun. Lett.*, vol. 13, no. 1, pp. 16–18, Jan. 2009.

[15] L.-C. Wang and A. Chen, "Optimal radio resource partition for joint contention- and connection-oriented multichannel access in OFDMA systems," *IEEE Trans. Mobile Comput.*, vol. 8, no. 2, pp. 162–172, Feb. 2009.

[16] C.-F. Tsai, C.-J. Chang, F.-C. Ren, and C.-M. Yen, "Adaptive radio resource allocation for downlink oFDMA/SDMA systems with multimedia traffic," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1734–1743, May 2008.

[17] D. Bartolome, A. I. Perez-Neira, and C. Ibars, "Practical bit loading schemes for multi-antenna multi-user wireless OFDM systems," in *Proc. 38th Asilomar Conf. Signals Syst. Comput.*, Nov. 2004, vol. 1, pp. 1030–1034.

[18] J. Cai, X. Shen, and J. W. Mark, "Downlink resource management for packet transmission in OFDM wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1688–1703, Jul. 2005.

[19] H. Wang and L. Dittmann, "Downlink resource management for QoS scheduling in IEEE 802.16 WiMAX networks," *Comput. Commun.*, vol. 33, no. 8, pp. 940–953, May 2010.

[20] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 66–71, Jan. 2009.

[21] C.-M. Yen, C.-J. Chang, and L.-C. Wang, "A utility-based TMCR scheduling scheme for downlink MIMO/OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 59, no. 8, pp. 4115–4150, Oct. 2010.

[22] A. J. Goldsmith and S. G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[23] C. T. Lin and C. S. George Lee, *Neural Fuzzy Systems*. Englewood Cliffs, NJ: Prentice–Hall, 1996.

[24] K.-R. Lo, C.-J. Chang, C. Chang, and C. B. Shung, "A QoS-guaranteed fuzzy channel allocation controller for hierarchical cellular systems," *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1588–1598, Sep. 2000.

[25] J. Ye, X. Shen, and J. W. Mark, "Call admission control in wideband CDMA cellular networks by using fuzzy logic," *IEEE Trans. Mobile Comput.*, vol. 4, no. 2, pp. 129–141, Mar./Apr. 2005.

[26] M. K. Tsay, Z. S. Lee, and C. H. Liao, "Fuzzy power control for downlink CDMA-based LMDS network," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3917–3921, Nov. 2008.

[27] C. L. Chen, J. W. Lee, C. Y. Wu, and Y. H. Kuo, "Fairness and QoS guarantees of WiMAX OFDMA scheduling with fuzzy controls," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, no. 6, pp. 1–14, Jan. 2009.

[28] N. S. Jayant and P. Noll, *Digital Coding for Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice–Hall, 1984.

[29] H. Yaghoobi, "Scalable OFDMA physical layer in IEEE 802.16 WirelessMAN," *Intel Technol. J.*, vol. 8, no. 3, 2004.

[30] "3GPP TR 25.892: Feasibility Study for OFDM for UTRAN Enhancement—Third-Generation Partnership Project," Eur. Telecommun. Std. Inst., Sophia-Antipolis, France, Tech. Rep., 2004.

[31] *Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS*, UMTS Std. 30.03, 1998.

[32] Voice Over IP—Per Call Bandwidth Consumption, CISCO Tech. Notes, Document ID 7934.

[33] Z. Diao, D. Shen, and V. O. K. Li, "An adaptive packet scheduling algorithm in OFDM systems with smart antennas," in *Proc. PIMRC*, 2005, pp. 2151–2155.

[34] "Wimax System Evaluation Methodology," WiMAX Forum, San Diego, CA, Tech. Rep., Jan. 2007.

[35] Y. Guo, Y. Chen, Y.-K. Wang, H. Li, M. M. Hannuksela, and M. Gabbouj, "Error resilient coding and error concealment in scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 6, pp. 781–795, Jun. 2009.

[36] H. Chen, H. C. B. Chan, V. C. M. Leung, and J. Zhang, "Cross-Layer enhanced uplink packet scheduling for multimedia traffic over MC-CDMA networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 986–992, Feb. 2010.

[37] B. Rong, Y. Qian, and K. Lu, "Downlink call admission control in multiservice WiMAX networks," in *Proc. IEEE ICC*, 2007, pp. 5082–5087.

[38] K. Suresh, I. S. Misra, and K. Saha, "Bandwidth and delay guaranteed call admission control scheme for QOS provisioning in IEEE 802.16e mobile WiMAX," in *Proc. IEEE GLOBECOM*, 2008, pp. 1–6.

Yao-Hsing Chung (S'08) received the B.S. and M.S. degrees in electrical engineering from the National Taipei University of Technology, Taipei, Taiwan, in 2004 and 2006, respectively. He is currently working toward the Ph.D. degree with the National Chiao Tung University, Hsinchu, Taiwan.

His research interests include wireless communication, communication protocol design, interference management, and voice over internet protocol.

Chung-Ju Chang (S'81–M'85–SM'94–F'05) was born in Taiwan in August 1950. He received the B.E. and M.E. degrees in electronics engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1972 and 1976, respectively, and the Ph.D. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1985.

From 1976 to 1988, he was with the Telecommunication Laboratories, Directorate General of Telecommunications, Ministry of Communications, Taiwan, as a Design Engineer, Supervisor, Project Manager, and then Division Director. He also acted as a Science and Technical Advisor for the Minister of the Ministry of Communications from 1987 to 1989. In 1988, he joined the Faculty of the Department of Communication Engineering, College of Electrical Engineering and Computer Science, National Chiao Tung University, as an Associate Professor. He has been a Professor since 1993 and a Chair Professor since 2009. He was the Director of the Institute of Communication Engineering from August 1993 to July 1995, the Chairman of the Department of Communication Engineering from August 1999 to July 2001, and the Dean of the Research and Development Office from August 2002 to July 2004. In addition, he was an Advisor for the Ministry of Education to promote the education of communication science and technologies for colleges and universities in Taiwan during 1995–1999. He is acting as a Committee Member of the Telecommunication Deliberate Body, Taiwan. His research interests include performance evaluation, radio resource management for wireless communication networks, and traffic control for broadband networks.

Dr. Chang is a member of the Chinese Institute of Engineers and the Chinese Institute of Electrical Engineers. He has served as Editor for the IEEE Communications Magazine and Associate Editor for the IEEE Transactions on Vehicular Technology.