Adaptive Radio Resource Allocation for Downlink OFDMA/SDMA Systems with Multimedia Traffic

Chun-Fan Tsai, Chung-Ju Chang, *Fellow, IEEE,* Fang-Ching Ren, *Member, IEEE,* and Chin-Ming Yen, *Student Member, IEEE*

***Abstract*— This paper proposes an adaptive radio resource allocation (ARRA) algorithm for downlink OFDMA/SDMA systems with multimedia traffic. Considering multiple service classes and diverse QoS requirements of multimedia traffic, the ARRA algorithm is designed with the goal to maximize spectrum efficiency and to fulfill quality of service (QoS) requirements. It is composed of two parts, a dynamic priority adjustment (DPA) scheme and a priority-based greedy (PBG) scheme. The DPA adopts a type of time-to-expiration to indicate the degree of user's urgency, and dynamically gives high priorities to urgent users. The PBG allocates the radio resource iteratively, based on a cost value, to maximize the system throughput while allocating enough resource to high-priority users. Simulation results show that the ARRA algorithm outperforms conventional algorithms in terms of system throughput and satisfaction extent of QoS requirements; it can sustain users' QoS requirements up till a traffic load of 0.8, while the conventional algorithms cannot guarantee QoS requirements after a traffic load of 0.3.**

***Index Terms*— Radio resource allocation, quality-of-service, scheduling, optimization, OFDM/SDMA.**

I. Introduction

ORTHOGONAL frequency division multiple access combined with space division multiple access (OFDM A/SDMA) can be an effective approach to support high-speed wireless communications. The OFDMA is based on OFDM (orthogonal frequency division multiplexing) and inherits its superiority of mitigating multipath fading and maximizing spectral efficiency (Nyquist rate). The SDMA uses a beamforming technique in a multiple-antenna system and multiplexes multiple users on the same subchannel to increase the system throughput.

For a multiuser OFDMA system, Jang and Lee proposed a linkgain-based resource allocation (LBRA) scheme and proved that the data rate of the system was maximized when each subcarrier was assigned to the user with the best channel gain [1]. However, this statement is not always true when SDMA is enabled in an OFDMA system. Instead, the system data rate is the largest when an optimal set of cochannel users, which depends on the spacial signature of each user, is selected for each subcarrier. But the algorithm for finding the optimal set of cochannel users is of a high computational complexity [2], [3]. Hence, when the channel state information (CSI) is available at the base station, a sophisticated and low-complexity radio resource allocation (RRA) scheme is needed for OFDMA/SDMA systems to properly exploit system diversity so that spectrum efficiency is maximized.

On the other hand, in a modern wireless system that supports multimedia traffic, quality-of-service (QoS) guarantee must be an essential design consideration for RRA algorithms. Song and Li [4] proposed a utility-based resource allocation and scheduling for OFDM-based wireless broadband networks. Although the utility-based resource allocation and scheduling is a good approach to deal with QoS issue, the utility function was not completely defined. Wong et al. [5] proposed an algorithm for the OFDMA system to minimize the total transmission power consumption while satisfying the QoS requirement, which was defined as the specified data transmission rate and bit error rate (BER). Efficient computational methods for reducing the complexity of the algorithm in [5] were presented in [6], [7]. Thoen et al. [8] investigated the same optimization problem as [5] but with the inclusion of SDMA. However, their proposed algorithm required the same cochannel users for each subcarrier and the optimization of the selection of cochannel users was not considered. Koutsopoulos and Tassiulas [9] first considered the rate maximization problem, where they tried to maximize the SIR (signal-to-interference ratio) for cochannel users, without the constraint of QoS requirement. They also proposed an algorithm for rate optimization under the same QoS requirement constraint as in [5]. However, in their proposal the power was fixed in each subcarrier and the SDMA was not enabled. Tsang and Cheng [10] challenged the performance of [8] and [9] and proposed an optimal solution for maximizing information capacity. Also, a multi-antenna multi-user maximum sum rate (MMSR) scheme was proposed for wireless OFDM systems in [11]. In the MMSR scheme, the user with the best channel quality is first selected for each subchannel. After that, users are added to a subchannel based on a *user clustering* procedure, where the main idea is to select the users with small spatial correlation in the same subchannel. The modulation order is then determined by a bit-removing algorithm. A heuristic approach was taken to reduce the high complexity of the MMSR scheme.

Generally, two kinds of optimization problem formulation for resource allocation in OFDM-based system can be found

Manuscript received December 11, 2006; revised July 18, 2007 and December 4, 2007; accepted January 18, 2008. The associate editor coordinating the review of this paper and approving it for publication was R. Murch. This work was supported by the National Science Council of Taiwan, ROC, under contract number NSC 95-2752-E-009-014 and the Ministry of Education (MOE) under ATU Program 95W803C.

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

Digital Object Identifier 10.1109/TWC.2008.060994.

in the literature, namely (i) power minimization [5], [8] and (ii) rate maximization [10], [11]. The former tries to minimize the overall transmit power given constraints on users' data rate or BER, while the latter tries to maximize the system data rate with constraints on the total power budget and on the users' BER performance. However, the fairness issue in the above two problem formulations is ignored. A new formulation to maximize the data rate with the constraint in which the allocated rate for each user is proportional to a specified weight was proposed in [12]. Cai, Shen, and Mark [13] studied a resource management scheme for packet transmission in OFDM wireless communication systems, and they proposed a truncated generalized processor sharing (TGPS) scheduling for cross-layer resource allocation in the MAC layer and a power and subcarrier allocation algorithm in the physical layer. The GPS scheduling algorithm requires a predefined weight for each data flow, and the allocated resource for each user is based on these weights. However, how to obtain the predefined weights for GPS scheduling or the proportional rate constraint is not specified. Huang and Letaief [14] adopted a rate-based scheduling algorithm to provide heterogeneous rate assignment for different users but not for multimedia traffics. As a result, the rate-based scheduling would not favor the user with strict QoS requirement and thus would not be applicable to the system with multimedia traffic.

From these previous works, three observations can be induced. First, all above schemes can be considered as fixed-priority schemes. The resource is either allocated to guarantee a fixed number of transmission bits in each OFDMA symbol, or assigned according to predefined weights for a GPS scheduling or proportional rate constraint. Since the required resource is fixed in each OFDMA symbol, the time diversity is not well exploited and it results in throughput degradation. As shown in [7], the system throughput increases with the number of users due to multiuser diversity but then decreases when the number of users is further increased. Second, only the bit error rate (BER) and/or the minimum transmission rate were considered as QoS requirements in previous RRA algorithms. However, with the presence of multimedia traffic, the delay requirement should also be included. Most packets should be transmitted within their delay bound, otherwise they will be dropped. Also, the packet dropping ratio should be kept below a desired level. And third, most studies assumed that a subcarrier is used as the basic allocation unit and that each user always has data in its buffer. However, a subcarrier-based allocation is difficult to realize due to its high control-signaling overhead. A basic allocation unit in a practical OFDMA system (e.g. IEEE 802.16 [15]) is a subchannel, which is a set of subcarriers. Moreover, in realistic environments providing multimedia service, the traffic models should be taken into account when designing RRA algorithms.

The paper proposes an adaptive radio resource allocation (ARRA) algorithm for downlink OFDMA/ SDMA systems with multimedia traffic. This radio resource allocation work is mathematically formulated into an optimization problem with an objective to maximize the system throughput under four designed constraints. For spectrum efficiency and QoS satisfaction, the radio resource allocated to a user have an upper bound and a lower bound, which are the result

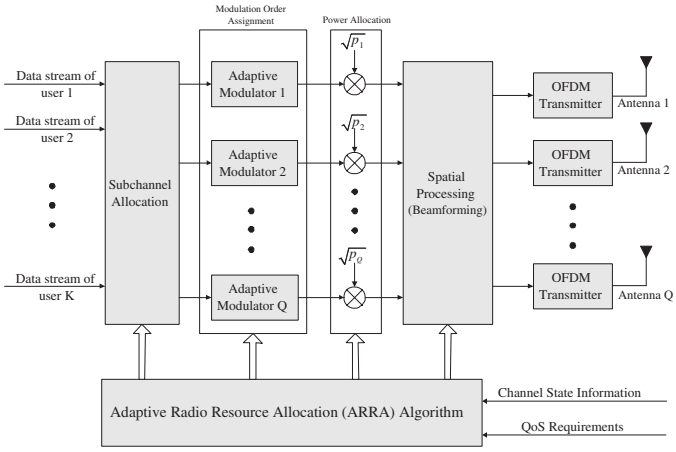

Fig. 1. The OFDMA/SDMA system.

of the *buffer occupation constraint* and the *QoS fulfillment constraint*, respectively. In addition, there are limitations on the system such as the total power and the number of users multiplexed on the same subchannel, and these limitations are represented by the *total system power constraint* and the *subchannel allocation constraint*. Moreover, the ARRA algorithm is composed of two parts to solve the optimization problem of radio resource allocation. The first part is a dynamic priority adjustment (DPA) scheme, where priorities of users are dynamically adjusted, based on a time-to-expiration (TTE) parameter and the radio resource required by each user, frame by frame. This can be considered as an MAC layer scheduling algorithm for determining the resource given to which user. With this scheme, it is believed that the ARRA algorithm can better attain the tradeoff between system throughput and QoS requirement than the schemes with a fixed priority. The second part is a low-complexity resource allocation scheme, called priority based greedy (PBG) scheme. The intention of the PBG scheme is to maximize the total system throughput under the four 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. Simulation results show that the ARRA algorithm can achieve the system throughput larger than conventional algorithms and guarantee QoS requirements until a rather higher traffic load.

The remainder of the paper is organized as follows. The system model of the considered OFDMA/SDMA system is introduced in section II. Section III presents the details of the proposed ARRA algorithm. Section IV discusses the performance of the ARRA algorithm. Finally, conclusions are given in section V.

II. System Model

A. OFDMA/SDMA System

The downlink OFDMA/SDMA system with the ARRA algorithm is shown in Fig. 1, where data streams for K single-antenna mobile stations are transmitted from the base station which is equipped with N subchannels and Q transmit antennas. A set of OFDM subcarriers forms an OFDMA subchannel. The mapping from subcarriers is assumed to be done regularly for ease of implementation [15], [16]. Also,

it has been shown that grouping of adjacent subcarriers will result in a highest multiuser diversity [16], which maximizes the system throughput. Hence, in this paper, a subchannel is assumed to have b adjacent subcarriers, and it is the basic unit for resource allocation and adaptive modulation. The time axis is divided into *frames* with fixed length, and each frame includes L symbols for OFDMA downlink transmission. The ARRA algorithm is executed at the beginning of every *frame* to properly allocate the radio resource to all users according to their queue state, CSI, and QoS requirements. Whenever a user attains the allocation, the ARRA algorithm continues the same allocation to the user in the following symbols of the frame until there are no more bits in the user's buffer to lessen the control signaling burden. Note that the radio resource allocation includes the subchannel allocation, modulation order assignment, power allocation, and beamforming control.

Let Ψ_n denote the set of the subcarriers in subchannel n and $\mathcal{K}_n^{(\ell)}$ denote the set of users that are multiplexed on subchannel n using the SDMA beamforming technology for the ℓth OFDMA symbol. Then, for the system under consideration, the transmit symbol vector in subcarrier i of subchannel n for the ℓth OFDMA symbol in a frame, denoted by $\mathbf{S}_i^{(\ell)}$, $1 \le \ell \le L$, is expressed as,

$$\mathbf{S}_i^{(\ell)} = \sum_{k \in \mathcal{K}_n^{(\ell)}} \sqrt{\xi_{k,i}^{(\ell)}} d_{k,i}^{(\ell)} \mathbf{w}_{k,i}^{(\ell)}, \qquad i \in \Psi_n, \tag{1}$$

where $\xi_{k,i}^{(\ell)}$ is the allocated power, $d_{k,i}^{(\ell)}$ is the data symbol, and $\mathbf{w}_{k,i}^{(\ell)}$ is a $Q \times 1$ beamforming vector, for user k in subcarrier i at the ℓth OFDMA symbol. Please note that a normalized QAM modulation is used such that the data symbol has unitary mean energy.

Assume that the coherent time of wireless channel is larger than the frame duration. Hence the channel is considered as fixed in a frame duration. Also, assume a perfect downlink CSI estimation for each user. Let $\mathbf{h}_{k,i}$ be a $1 \times Q$ vector denoting the frequency domain channel gain from the base station to user k on subcarrier i. Note that $\mathbf{h}_{k,i}$ is not a function of ℓ since the channel is assumed to be fixed within a frame. For the sake of simplicity and acceptable performance, a zero-force (ZF) transmit beamforming scheme in [2], [10], [11] is used. The performance of ZF beamforming is equivalent to that of the minimum mean square error (MMSE) beamforming for a low number of cochannel users or a high SNR [8]. Since a RRA scheme usually considers to allocate resource to users with good channel qualities, the SNR is usually high and thus the ZF beamforming can achieve a close performance to the MMSE beamforming. With the ZF transmit beamforming where users are orthogonal in space domain, the received signal of user k in subcarrier i for the ℓth OFDMA symbol, denoted by $Y_{k,i}^{(\ell)}$, is given by,

$$Y_{k,i}^{(\ell)} = \mathbf{h}_{k,i} \mathbf{w}_{k,i}^{(\ell)} \sqrt{\xi_{k,i}^{(\ell)}} d_{k,i}^{(\ell)} + Z_{k,i}^{(\ell)}, \tag{2}$$

where $Z_{k,i}^{(\ell)}$ is the thermal noise on user k in subcarrier i and is assumed to be in complex Guassian distribution with zero mean and variance σ^2. Then, the received SNR of user k in subcarrier i for the ℓth OFDMA symbol, denoted by $SNR_{k,i}^{(\ell)}$, can be obtained by,

$$SNR_{k,i}^{(\ell)} = \frac{\xi_{k,i}^{(\ell)} \left| \mathbf{h}_{k,i} \mathbf{w}_{k,i}^{(\ell)} \right|^2}{\sigma^2}. \tag{3}$$

The received SNR is affected by the beamforming vector. If the user k has a high spatial correlation, the term $\mathbf{h}_{k,i} \mathbf{w}_{k,i}^{(\ell)}$ will be small and a poor received SNR will be the result. Note that the ARRA algorithm will select those cochannel users such that their effective linkgains are large enough.

B. Power Allocation

The allocated power to user k on subcarrier i for the ℓth OFDMA symbol in a frame, $\xi_{k,i}^{(\ell)}$, $1 \le \ell \le L$, is determined by the minimum required SNR of user k, which can be obtained from the QoS requirement of BER, BER_k^*, and the modulation scheme of user k. If user k adopts M-QAM modulation, the minimum required SNR, SNR_k^*, is given by [17],

$$SNR_k^* = -\frac{\ln(5\, BER_k^*)}{1.5}(M-1). \tag{4}$$

Note that (4) represents an approximation on the required SNR, whereas the exact SNR thresholds are given in [18], [19]. The allocated power, $\xi_{k,i}^{(\ell)}$, is set to the value such that the $SNR_{k,i}^{(\ell)}$ in (3) is equal to the SNR_k^* in (4), hence it can be obtained by,

$$\xi_{k,i}^{(\ell)} = \frac{-\ln(5\, BER_k^*)(M-1)}{1.5} \frac{\sigma^2}{\left| \mathbf{h}_{k,i} \mathbf{w}_{k,i}^{(\ell)} \right|^2}. \tag{5}$$

Consequently, the power allocated to user k on subchannel n for the ℓth OFDMA symbol in a frame, denoted by $p_{k,n}^{(\ell)}$, $1 \le \ell \le L$, can be calculated as,

$$p_{k,n}^{(\ell)} = \sum_{i \in \Psi_n} \xi_{k,i}^{(\ell)}. \tag{6}$$

In other words, the power allocated to a user should be sufficient to guarantee the BER requirement if the user is selected by the ARRA algorithm.

C. Service Classes

The OFDMA/SDMA system is assumed to support three classes of services, real-time (RT), non-real-time (NRT), and best effort (BE), which are with differentiated QoS requirements. For RT services, the QoS requirements consider BER, maximum packet delay tolerance, and maximum packet dropping ratio. For NRT services, the QoS requirements are BER and minimum required transmission rate. For BE services, only BER is included in the QoS requirement. Each mobile station (user) belongs to one kind of service class, and a traffic model is associated with the user. Denote these QoS requirements of BER, minimum required transmission rate, maximum packet delay tolerance, and maximum packet dropping ratio by BER_k^*, R_k^*, D_k^*, and $P_{D,k}^*$ respectively. Also, four kinds of traffic types are assumed in this system, voice and video traffic of RT service, HTTP traffic of NRT service, and FTP traffic of BE service.

The OFDMA/SDMA system provides one individual queue for each traffic type of downlink user at the base station. Arriving packets for each user are stored in their own individual queue in a first-in first-out manner. Packets of RT services will be dropped if the packet delay exceeds the maximum packet delay tolerance, while packets of NRT services or BE services are allowed to queue without being dropped if the buffer occupancy is not overflowed. Retransmission due to erroneous transmission of packets is not considered in this paper.

III. Adaptive Radio Resource Allocation

The adaptive radio resource allocation (ARRA) algorithm is designed to determine an optimal assignment such that the total system throughput is maximized while each user's QoS requirements are satisfied. Define $x_{k,n}^{(\ell)}$ as the assignment variable of modulation order for user k on subchannel n for the ℓth OFDMA symbol, where $x_{k,n}^{(\ell)} \in \{0,1,2,3\}$, $1 \le k \le K$, $1 \le n \le N$, and $1 \le \ell \le L$. If $x_{k,n}^{(\ell)} = 0$, it denotes that the data for user k is not transmitted on subchannel n at the ℓth OFDMA symbol. If $x_{k,n}^{(\ell)} = 1, 2,$ or 3, it means that the data for user k is transmitted on this subchannel using modulation scheme of QPSK, 16-QAM, or 64-QAM, respectively, at the ℓth OFDMA symbol. Denote the assignment vector $\mathbf{x}^{(\ell)} \equiv \left[x_{1,1}^{(\ell)}, \cdots, x_{1,N}^{(\ell)}, \cdots, x_{k,1}^{(\ell)}, \cdots, x_{k,N}^{(\ell)}, \cdots, x_{K,1}^{(\ell)}, \cdots, x_{K,N}^{(\ell)}\right]^T$ the solution of the ARRA algorithm for the ℓth OFDMA symbol. The throughput of user k is defined as the allocated transmission bits to user k in this frame, denoted by R_k, and can be calculated from $\mathbf{x}^{(\ell)}$, $1 \le \ell \le L$, by

$$R_k = R_k(\mathbf{x}^{(1)} \cdots \mathbf{x}^{(\ell)} \cdots \mathbf{x}^{(L)}) = \sum_{\ell=1}^{L} \sum_{n=1}^{N} q \cdot x_{k,n}^{(\ell)}, \quad (7)$$

where $q = 2 \times b$ is the number of transmission bits with the basic QPSK modulation over b subcarriers in one subchannel. Also, the selected user set in subchannel n at the ℓth OFDMA symbol, $\mathcal{K}_n^{(\ell)}$, can be obtained from $\mathbf{x}^{(\ell)}$ by,

$$\mathcal{K}_n^{(\ell)} = \mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)}) = \left\{k | x_{k,n}^{(\ell)} > 0, 1 \le k \le K\right\}. \quad (8)$$

It is evident from (5) and (6) that the allocated power, $p_{k,n}^{(\ell)}$, is a function of BER_k^* and $\mathbf{x}^{(\ell)}$. Hence, if needed, $p_{k,n}^{(\ell)}$ will be denoted by $p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)})$ in the following.

The ARRA algorithm formulates the RRA problem as an optimization problem given by,

$$(\mathbf{x}^{*(1)} \cdots \mathbf{x}^{*(L)}) = \arg \max_{\mathbf{x}^{(1)} \cdots \mathbf{x}^{(L)}} \sum_{k=1}^{K} R_k(\mathbf{x}^{(1)} \cdots \mathbf{x}^{(L)})$$

subject to the following constraints:

$$\begin{aligned}
&\text{(i)} \quad |\mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)})| \le Q, \ \forall n, \ell, \\
&\text{(ii)} \quad \sum_{n=1}^{N} \sum_{k=1}^{K} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}) \le P_T, \ \forall \ell, \\
&\text{(iii)} \quad R_k \le \left\lceil R_k^B / q \right\rceil \cdot q, \ \forall k, \\
&\text{(iv)} \quad R_k \ge \widehat{R}_k, \ \forall k,
\end{aligned} \quad (9)$$

where P_T is the limit of the total power allocation constraint for every OFDMA symbol of a frame, R_k^B is the user's buffer occupancy at the beginning of this frame, and $\widehat{R}_k$ is the priority value of user k. Constraint (i) is the *subchannel allocation constraint* because a subchannel can be allocated to at most Q users for each OFDMA symbol in the OFDMA/SDMA system. Constraint (ii) is the *total system power constraint* because the total power allocation for downlink data transmission at the base station should have a limitation for each OFDMA symbol. Constraint (iii) is the *buffer occupation constraint* for spectrum efficiency. The allocated transmission bits to user k in a frame, R_k, should not be larger than R_k^B (bits). The final constraint (iv), referred to as the *QoS fulfillment constraint*, is required to further satisfy the QoS requirement of maximum packet delay tolerance for RT users and the QoS requirement of minimum required transmission rate for NRT users. Here a priority value $\widehat{R}_k$ is set for each user k at each frame according to its QoS requirements and queue state. We define the $\widehat{R}_k$ as the minimum number of bits required to transmit at the current frame otherwise the user's QoS requirements cannot be fulfilled. The larger $\widehat{R}_k$, the higher the priority and the more the resource should be allocated to user k. Thus the R_k should have this constraint, which means that the user k should be allocated with at least $\widehat{R}_k$ bits at the current frame in order to satisfy its QoS requirements. Noticeably, the $\widehat{R}_k$ is dynamically adjusted frame by frame.

The ARRA algorithm also finds the optimal set of assignment vector for a frame, $(\mathbf{x}^{*(1)} \ldots \mathbf{x}^{*(L)})$, so that the total system throughput is maximized under the four system constraints. Here, the proposed ARRA algorithm adopts a reduced-complexity approach which is based on the greedy algorithm [6], [20]. The greedy approach has been shown that it can achieve a near-optimal solution with lower computational complexity. Therefore, the proposed ARRA algorithm contains two parts for solving the optimization problem given in (9). The first part is a dynamic priority adjustment (DPA) scheme and the second part is a priority-based greedy (PBG) scheme. The details are described in the following.

A. DPA Scheme

Here we introduce a *time-to-expiration* (TTE) parameter to indicate the urgency degree of a user at the current frame for the DPA scheme. For user k, we denote the TTE parameter and the number of residual bits of the head-of-line (HOL) packet by V_k and B_k, respectively. The smaller V_k, the more the degree of urgency is of user k. For users with RT service class, the V_k is intuitively given by,

$$V_k = D_k^* - D_k, \quad (10)$$

where D_k is the delay time from the arrival of the HOL packet of user k to the current frame, and the unit of both D_k and D_k^* is in frames. For users with NRT service class, the V_k is given by,

$$V_k = \left\lfloor \frac{B_k' + B_k}{R_k^*} - D_k' \right\rfloor, \quad (11)$$

where $\lfloor x \rfloor$ is the largest integer smaller than x, D_k' is the time duration while there is data buffered in the queue of user k before the current frame, B_k' is the total number of

transmission bits of user k in D'_k, and R^*_k is the minimum required transmission rate in a unit of bits per frame. The derivation of V_k in (11) of NRT user k comes from the inequality $(B_k + B'_k)/(V_k + D'_k) \geq R^*_k$, which means that the average rate should be greater than the minimum required transmission rate. An NRT user's HOL packet should complete its transmission within its TTE value, otherwise the rate requirement of the user is not satisfied. Finally, for users with BE service class, the $V_k = \infty$ since there is no delay or rate requirement for BE users.

Given V_k and B_k of user k, its priority value at the beginning of a frame, $\widehat{R}_k$, is defined as,

$$\widehat{R}_k = \begin{cases} 0, & \text{if } V_k = \infty \\ \left\lceil \frac{B_k}{q} \right\rceil \cdot q, & \text{if } V_k \leq V_{th} \\ \max\left(\left\lceil \frac{B_k}{V_k \cdot q} \right\rceil - \lceil \ln(V_k) \rceil, 0 \right) \cdot q, & \text{elsewise,} \end{cases} \tag{12}$$

where $\lceil x \rceil$ is the smallest integer larger than x and V_{th} is a threshold for V_k. If $V_k = \infty$, it is intuitive to set $\widehat{R}_k$ as zero. If V_k is below threshold V_{th}, it means that the degree of urgency of user k is very high such that user k should complete its transmission in this current frame, thus $\widehat{R}_k$ is set equal to $\left\lceil \frac{B_k}{q} \right\rceil \cdot q$. Otherwise, the design of $\widehat{R}_k$ is based on the average required transmission bits in remaining frames, B_k/V_k, added with a negative bias $(-\lceil \ln(V_k) \rceil)$. The negative bias reduces the priority of the delay-tolerable users, who have a large V_k, so that the system can give the transmission opportunity to other high-urgency users. Note that a user with a low priority could still be served by the base station if the channel quality of the user is good and if other users with higher priority have already been served. Hence, the delay-tolerable users can take advantage of the time diversity by transmitting only when its channel is good, thereby enhancing the system throughput. As for the threshold value V_{th}, it could be set to one if the resource is always enough to satisfy $R_k \geq \widehat{R}_k$. However, since the user might be at cell boundary, the V_{th} should be set to a larger value to guarantee the QoS requirement earlier.

B. PBG Scheme

The PBG scheme is designed to maximize the total system throughput. Here, we define an immediate cost as the increment of power by increasing one modulation order for a user on one subchannel in every successive iteration. If the immediate cost can be minimized, the total system throughput can be maximized.

The cost function of user k on subchannel n at the ℓth OFDMA symbol, denoted by $C^{(\ell)}_{k,n}$, is expressed as,

$$C^{(\ell)}_{k,n} = \begin{cases} \sum_{k \in \mathcal{K}^{(\ell)}_n(\mathbf{x}^{+(\ell)})} [p^{(\ell)}_{k,n}(BER^*_k, \mathbf{x}^{+(\ell)}) \\ \qquad - p^{(\ell)}_{k,n}(BER^*_k, \mathbf{x}^{(\ell)})], \\ \qquad \text{if } 0 \leq x^{(\ell)}_{k,n} \leq 2 \text{ and } |\mathcal{K}^{(\ell)}_n(\mathbf{x}^{+(\ell)})| \leq Q, \\ \infty, \quad \text{otherwise,} \end{cases} \tag{13}$$

where the term in the square bracket is the immediate cost , and $\mathbf{x}^{+(\ell)}$ is the assignment vector after the modulation of user k on subchannel n is increased by one, given the current $\mathbf{x}^{(\ell)}$. The $\mathbf{x}^{+(\ell)}$ will be the same as $\mathbf{x}^{(\ell)}$ except $x^{+(\ell)}_{k,n} = x^{(\ell)}_{k,n} + 1$. Note that the increase of the modulation order from zero to one means adding a new user to a subchannel. Since adding a new user to a subchannel requires recalculation of beamforming vectors, the required transmission power for maintaining the same modulation order for the users that are already in the subchannel is increased. The cost function, defined in (13), also includes the above increasing power. Hence the spatial correlation between the new user and the users that are already in the subchannel is also measured by the cost function.

At the beginning of a frame, the PBG scheme receives the information $\left\{ \widehat{R}_k, 1 \leq k \leq K \right\}$ from the result of the DPA scheme and initially sets the assignment variables, $x^{(\ell)}_{k,n}, \forall k, n$, the current used power, denoted by $P^{(\ell)}$, to be zero, and the set of free subchannels available for the ℓth OFDMA symbol, denoted by $\mathcal{N}^{(\ell)}_{free}$, to be $\mathcal{N}^{(\ell)}_{free} = \{n | 1 \leq n \leq N\}$, for all ℓ. The PBG scheme then finds the optimal assignment vectors, $(\mathbf{x}^{(1)} \ldots \mathbf{x}^{(L)})$, which is the solution of the resource allocation in a frame.

For detailed description, a pseudocode of the PBG scheme is given in Appendix I. The PBG scheme runs in an iterative process for symbol ℓ to find $\mathbf{x}^{(\ell)}$, $1 \leq l \leq L$. It constructs a candidate user set, denoted by Ω, and selects the optimal pair of user and subchannel, denoted by (k^*, n^*), which is defined as the highest priority user on the subchannel such that its cost value is the smallest. The Ω contains the backlogged users with the highest priority, and the (k^*, n^*) is obtained by choosing the user in Ω and the subchannel in $\mathcal{N}^{(\ell)}_{free}$ such that the cost value, $C^{(\ell)}_{k^*,n^*}$, is minimal. If the power budget in the ℓth OFDMA symbol is still sufficient for increasing the modulation order for user k^* on subchannel n^*, then the modulation order of the pair (k^*, n^*) is increased by one, i.e. $x^{(\ell)}_{k^*,n^*} = x^{(\ell)}_{k^*,n^*} + 1$. In each iteration, if q bits are allocated to the selected user k^*, then the queue length of user k^*, $R^B_{k^*}$, will be decreased by q and the used power for the ℓth OFDMA symbol, $P^{(\ell)}$, is increased by the minimum cost. Also, the priority value of user k^* will be decreased by q or equal to zero, i.e. $\widehat{R}_{k^*} = \max(\widehat{R}_{k^*} - q, 0)$.

To reduce the complexity of the ARRA algorithm and the control signaling overhead of the system, the PBG scheme continues the same allocation as the one for the ℓth OFDMA symbol to its next consecutive symbols to form a time burst transmission. For each subchannel in $\mathcal{N}^{(\ell)}_{\text{free}}$, it pre-assigns the same assignment variables for the ℓth OFDMA symbol to the next successive ones, i.e. $x^{(i)}_{k,n} = x^{(\ell)}_{k,n}$, for $l + 1 \leq i \leq L$, and allocates $x^{(\ell)}_{k,n} \cdot q$ bits to user k for the ith symbol for all k, $l + 1 \leq i \leq L$. This step will continue until the *buffer occupation constraint* is violated, i.e. $\lceil R^B_k / q \rceil \cdot q < x^{(\ell)}_{k,n} \cdot q$. The assignment variables, priority value, queue length, and used power are updated if this step is executed. Also, subchannel n should be removed from the set of $\mathcal{N}^{(i)}_{\text{free}}$ since it has been allocated at the ith symbol.

IV. Simulation Results and Discussions

A. Simulation Environment

The downlink OFDMA/SDMA system environment is set to be compatible to the IEEE 802.16 standard [15] in the simulations, where parameters are listed in Table I and scalable

TABLE I
OFDMA/SDMA SYSTEM PARAMETERS

Parameters	Values
Cell size	1600m
Number of antenna at base station (Q)	3
Frame duration	2ms
System bandwidth	5 MHz
FFT size	512
Subcarrier frequency spacing	11.16 kHz
OFDMA symbol duration	100.8 μsec
Number of data subcarriers	384
Number of subchannels (N)	8
Number of data subcarriers per subchannel (b)	48
Number of OFDMA symbol for downlink transmission per frame (L)	8
Power allocation to data transmission (P_T)	43.10 dBm
Thermal noise density	-174 dBm/Hz

parameters in the physical layer are configured according to the suggested values in [21]. Also, the path loss model is modeled as $128.1 + 37.6 \log R$ dB, where R, in unit of kilometers, is the distance between the base station and the user [22]. The log-normal shadowing is assumed to be with zero mean and standard deviation of 8 dB, and the multipath channel for each antenna has six taps of Rayleigh-faded paths with an exponential power delay profile. Various users would have independent channels but with the same statistics. The threshold V_{th} for V_k is set to be three.

There are four traffic types. The first one is the voice traffic of RT service. Each voice traffic is modeled as an ON-OFF model, in which lengths of ON period and OFF period follow an exponential distribution with means 1.0 and 1.5 seconds [23], respectively. The second one is the streaming video traffic of RT service. Each frame of video data is assumed to arrive at a regular interval of 100ms, each frame is decomposed into eight slices (packets), and the size of a packet is distributed in a truncated Pareto distribution [22]. Also, there are delay intervals between two consecutive packets of a frame which denote the encoding delay at the video encoder. These intervals are modeled by a truncated Pareto distribution. The third one is the HTTP traffic [22] of NRT service, where the behavior of web browsing is modeled. Thus, the traffic of the HTTP user is modeled as a sequence of page downloads, and each page download is modeled as a sequence of packet arrivals. The interval between two consecutive page downloads, representing the reading time in web browsing, is distributed in an exponential distribution. For detailed parameters of video and HTTP traffic models please refer to [22]. The last traffic type is the FTP traffic [22] of BE service. Each FTP user data is modeled as a sequence of file downloads. The size of a file is distributed in a truncated lognormal distribution with mean 2M bytes, standard deviation 0.722M bytes, and a maximum value 5M bytes. In addition, the interval between files is distributed in an exponential distribution with mean 180 seconds. The QoS requirements of each traffic type are listed in Table II [24]-[25].

TABLE II
THE QOS REQUIREMENT OF EACH TRAFFIC TYPE

	Voice (RT)	Video (RT)	HTTP (NRT)	FTP (BE)
Required BER	10^{-3}	10^{-4}	10^{-6}	10^{-6}
Maximum Packet Delay Tolerance	40 ms	10 ms	N/A	N/A
Maximum Packet Dropping Ratio	1%	1%	N/A	N/A
Minimum Required Transmission Rate	N/A	N/A	100 kbps	N/A

N/A : Not Applicable.

Next, the ARRA algorithm will be compared to three conventional RRA schemes with some modifications described in the following. (i) *Linkgain-based resource allocation (LBRA) scheme* [1]: The original scheme does not consider multiple antenna. Here it is modified so that the best Q users are selected to fit the architecture of multiple antennas. (ii) *Multi-antenna Multi-user Maximum Sum Rate (MMSR) scheme* [11]. (iii) *Truncated Generalized Processor Sharing (TGPS) scheme* [13]: Here we assume that the modulation scheme for TGPS is fixed at 16-QAM since the performance of TGPS is the best while using this modulation level. The predefined weight for TGPS is set to 10, 5, and 1 for RT, NRT, and BE services, respectively.

B. Performance Evaluation

In the simulations, the number of users is increased from 40 to 600, and the number of users in each traffic type is assumed to be the same. We define the traffic load of the system as the ratio of the total average data rate of users over the maximum system transmission rate. The maximum system transmission rate is achieved when Q users are multiplexed for each subchannel and the highest modulation order is used for all users. This is equal to 27.648 Mbps in this simulation environment. Also, the average date rate of each voice, video, HTTP, or FTP arrival user is equal to 5.2 kbps, 64 kbps, 14.5 kbps, or 88.9 kbps, respectively. Thus, the traffic load varies from 0.06 to 0.93 as the number of users varies from 40 to 600.

The following performance measures are investigated in the simulations: (i) system throughput, (ii) packet dropping ratio of RT users, (iii) mean packet delay of RT users, (iv) average transmission rate of NRT users, (v) guaranteed ratio of NRT users, defined as the ratio of the number of NRT users, whose average transmission rates are larger than the minimum required transmission rate, over the total number of NRT users, and (vi) average transmission rate of BE users. Finally, the complexity analysis is also discussed.

Fig. 2 shows the system throughput versus the traffic load. It can be found that the system throughput of the ARRA scheme performs the best and can reach up to 24 Mbps,

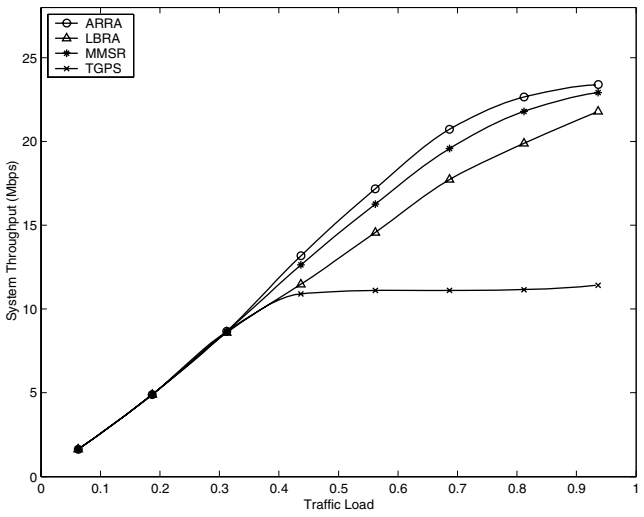

Fig. 2. System throughput.

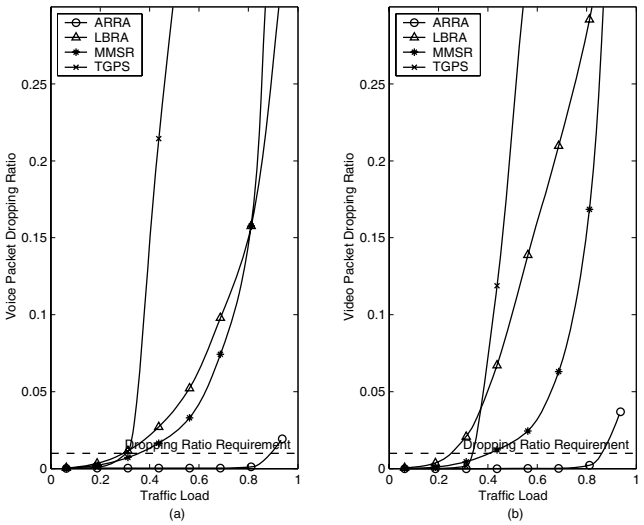

Fig. 3. (a) Packet dropping ratio of voice users; (b) packet dropping ratio of video users.

which is very closed to the maximum system throughput of 27.648 Mbps. The reasons for this are: the ARRA algorithm improves the system throughput by taking multiuser diversity and space domain correlation between users into account. The system throughput of the MMSR scheme is near to that of the ARRA scheme because both of there two schemes take throughput maximization as one of the design objectives. The system throughput of the LBRA scheme is less than that of the ARRA scheme due to the fact that the optimal user grouping in the space domain is not considered in the LBRA scheme. The system throughput of the TGPS scheme is the smallest since the TGPS uses a simplified algorithm for subchannel allocation and the multiuser diversity is not well exploited.

Figs. 3 (a) and 3 (b) depict the packet dropping ratios of voice users and video users, respectively, including the QoS requirement $P^*_{D,k}$ in dotted line. These figures show that the voice and video packet dropping ratios of the ARRA algorithm are almost zero until the traffic load becomes greater than 0.8, while those of the other conventional schemes increase rapidly

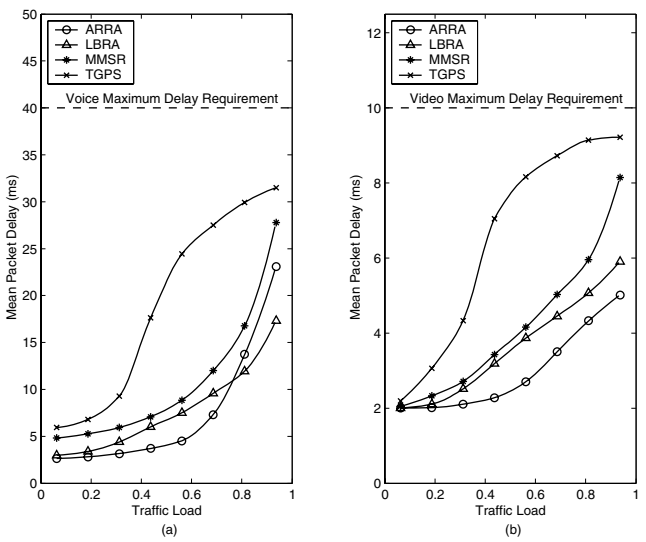

Fig. 4. (a) Mean packet delay of voice users; (b) mean packet delay of video users.

with the traffic load and violate the requirement of the packet dropping ratio at a traffic load of 0.3 or even 0.2. The reason is that the LBRA or MMSR scheme does not consider the QoS requirement of maximum packet delay tolerance for RT users. As for the TGPS scheme, since its maximum capacity is small, thus the packet dropping ratio becomes large at a high traffic load even though it gives high weights on RT users. On the other hand, the ARRA algorithm promotes the RT users with the larger packet delay as more urgent users and gives them higher priority. Therefore, the packet dropping ratio can be small and the delay requirement of RT users can be satisfied.

Figs. 4 (a) and 4 (b) show the mean packet delay of voice users and video users, respectively, where the maximum packet delay requirement D^*_k is also included. For all the algorithms, the mean packet delays are lower than the delay requirement. It can also be found that, in most cases, the delay of the ARRA algorithm is lower than those of TGPS, LBRA, or MMSR schemes. The reason for this is that the ARRA algorithm achieves the largest system throughput and gives the priority of RT users higher than other users. The mean packet delay for voice users of the LBRA scheme is less than that of the ARRA algorithm when the traffic load is higher than 0.7. However, at that traffic load, the packet dropping ratio for voice users of the LBRA scheme is much higher than that of the ARRA algorithm and violates the QoS requirement of maximum packet dropping ratio $P^*_{D,k}$. Consequently, we can conclude from Fig. 3 and Fig. 4 that the ARRA algorithm outperforms the conventional methods under the QoS requirements of RT users.

Figs. 5 (a) and 5 (b) illustrate the average transmission rate of HTTP users and the guaranteed ratio of HTTP users, respectively. For the ARRA algorithm, the average transmission rate decreases as the traffic load increases, but the minimum required transmission rate for NRT users is guaranteed. The ARRA algorithm guarantees the minimum transmission rate of each NRT user by giving high priority to the NRT users with a transmission rate lower than the minimum required transmission rate. For the same reason, the guaranteed ratio

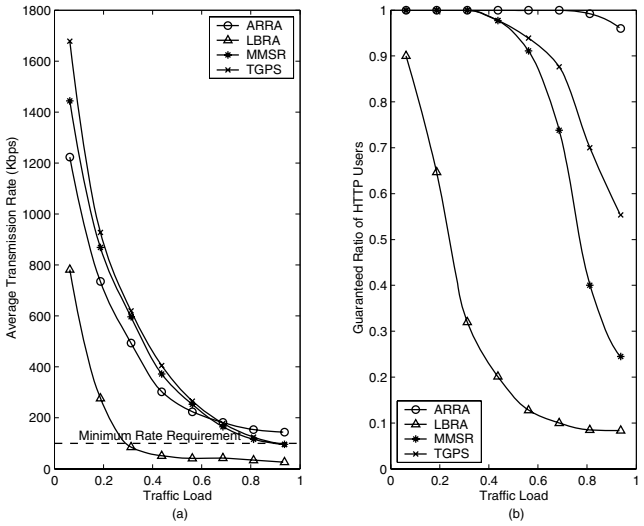

Fig. 5. (a) Average transmission rate of HTTP users; (b) guaranteed ratio of HTTP users.

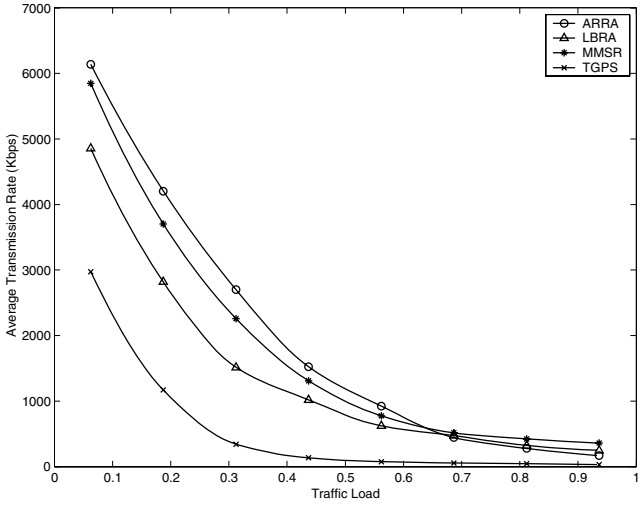

Fig. 6. Average transmission rate of FTP users.

of HTTP users is almost 100% in the ARRA algorithm when the traffic load is not too high, and is still larger than 95% when the traffic load is 0.9. Although the average transmission rate of the MMSR or TGPS scheme is higher than that in the ARRA algorithm, the guaranteed ratio of HTTP users drops earlier than the ARRA algorithm when the traffic load becomes larger. For example, when the traffic load is 0.8, the guaranteed ratio of the ARRA algorithm is 99% while those of the TGPS scheme and the MMSR scheme are only 70% and 40%, respectively. The LBRA scheme has the lowest guaranteed ratio of HTTP users since it only guarantees the transmission rate of users with good channel quality. The ARRA algorithm can have the average transmission rate of all NRT users satisfactory and guarantee each NRT user with a minimum transmission rate.

Fig. 6 shows the average transmission rate of FTP users with BE service. Although it is not required to guarantee the minimum transmission rate for BE users, the ARRA algorithm still gets the transmission rate for BE service higher than other schemes when the traffic load is less than 0.7. This is because

TABLE III
COMPLEXITY COMPARISON

Algorithm	Complexity
ARRA	$O(LKQN^2)$
LBRA	$O(LKQN)$
MMSR	$O(LKN(K+NQ))$
TGPS	$O(LNQ(K+N))$
Exhaustive Search	$O(L4^{KN})$

the goal of throughput maximization in the ARRA algorithm makes the transmission rate of each user as large as possible. However, its transmission rate is lower than that of the LBRA or MMSR scheme at a higher traffic load dut to the fact that the ARRA algorithm has to guarantee the QoS requirements for other high priority service classes. We consider this to be a worthwhile tradeoff since the ARRA algorithm performs much better than the other schemes in the achievement of high system throughput and in satisfying QoS requirements, as shown in previous figures.

Furthermore, the ARRA algorithm can efficiently solve the multi-dimensional (space, time, and frequency) RRA problem. The DPA scheme sequentially sets the priority for allusers; the complexity is $O(K)$. The PBG scheme finds a best pair of user and subchannel from K users and N subchannel; its complexity is $O(NK)$ in each iteration; and one can easily verify that the number of iterations is bounded by $3NQ$ since the system reaches its maximum capacity after this number of iterations. The complexity of finding $\mathbf{x}^{(\ell)}$ is $O(KQN^2)$ in the worse case. Therefore, the worse-case computational complexity for the PBG scheme is $O(LKQN^2)$. Consequently, the overall complexity of the ARRA algorithm is $O(K)+O(LKQN^2) = O(LKQN^2)$. However, in practical the PBG scheme continues the same allocation for several next OFDMA symbols, thus the complexity would be greatly reduced by almost L times. On the other hand, the MMSR (TGPS) scheme has a computational complexity in the order of $O(LKN(K+NQ))(O(LNQ(K+N)))$. The LBRA scheme has the complexity in the order of $O(LKQN)$.

Table III shows the computational complexity of the ARRA algorithm, the three conventional algorithms, and the exhaustive search method. The ARRA algorithm, the MMSR scheme, and the TGPS scheme have the similar computational complexity; the LBRA scheme gets the simplest complexity. Although the LBRA scheme has the lowest complexity among the algorithms, it performs quite poorly in satisfying of QoS requirements. It can be concluded that the ARRA algorithm outperforms the conventional schemes with acceptable complexity.

V. CONCLUSIONS

In this paper, an adaptive radio resource allocation (ARRA) algorithm is proposed for downlink OFDMA/SDMA systems with multimedia traffic, where the radio resource allocation includes subchannel allocation, modulation order assignment, power allocation, and beamforming. The goal of the ARRA algorithm are to fulfill QoS requirements for users and to maximize spectrum efficiency for system, while considering

multiple service classes, such as RT, NRT and BE services. The proposed ARRA algorithm contains a DPA scheme to dynamically adjust the priority of users frame by frame and a PBG scheme to efficiently allocate the resource based on a cost value.

Perfect CSI estimation is assumed to support the proposed ARRA algorithm. Generally speaking, a user would be allocated radio resource on better channels to have higher throughput. In practical, the CSI feedback complexity can be reduced by using a formulated codebook design [26], [27], and a user has to report the CSI for only the top 3 channels instead of all channels state information. Therefore, it will not be an overhead burden to the proposed ARRA algorithm.

Simulation results show that the ARRA algorithm outperforms the conventional algorithms in terms of system throughput and the extent to which QoS requirements are satisfied. The ARRA algorithm can sustain users' QoS requirements up to the traffic load of 0.8 while the conventional algorithms cannot guarantee users' QoS requirements at the traffic load greater than 0.3. Also, the system throughput of the ARRA algorithm is larger than conventional algorithms. This is because the ARRA algorithm can dynamically adjust the priority values of users which indicate the users' urgency degree, frame by frame. This makes NRT users with a low average transmission rate and RT users with a large packet delay, at a given time frame, obtain the resource earlier. In addition, the ARRA algorithm takes throughput maximization as its objective when there are no urgent users. As a result, the ARRA algorithm can achieve a higher system throughput under the fulfillment of users' QoS requirements.

APPENDIX I
PSEUDOCODE OF THE PBG SCHEME

- **[PBG Scheme]**

Receive $\left\{\widehat{R}_k | 1 \le k \le K\right\}$.
Set $\mathbf{x}^{(\ell)} = \mathbf{0}$, $P^{(\ell)} = 0$, *for* $1 \le \ell \le L$
Set $\mathcal{N}_{free}^{(\ell)} = \{n | 1 \le n \le N\}$, *for* $1 \le \ell \le L$.
for $\ell = 1 : L$ **do**
  [Resource allocation for symbol ℓ***]***
  if $|\mathcal{N}_{free}^{(\ell)}| = 0$ **then**
    Set $\ell = \ell + 1$.
    go to: ***Resource allocation for symbol*** ℓ.
  end if
  Set $\Omega_K = \left\{k | R_k^B > 0, 1 \le k \le K\right\}$.
  while $|\Omega_K| > 0$
    Set $\widehat{R}_{max} = \max_{k \in \Omega_K} \widehat{R}_k$.
    Set $\Omega = \left\{k | \widehat{R}_k = \widehat{R}_{max}, k \in \Omega_K\right\}$.
    Find $(k^*, n^*) = \arg\min_{k \in \Omega, n \in \mathcal{N}_{free}^{(\ell)}} C_{k,n}^{(\ell)}$.
    if $C_{k^*,n^*}^{(\ell)} + P^{(\ell)} > P_T$, **then**
      Set $\ell = \ell + 1$.
      go to: ***Resource allocation for symbol*** ℓ.
    end if
    Set $x_{k^*,n^*}^{(\ell)} = x_{k^*,n^*}^{(\ell)} + 1$.
    Set $R_{k^*}^B = \max(R_{k^*}^B - q, 0)$.
    Set $P^{(\ell)} = P^{(\ell)} + C_{k^*,n^*}^{(\ell)}$.
    Set $\widehat{R}_{k^*} = \max(\widehat{R}_{k^*} - q, 0)$.
    if $R_{k^*}^B = 0$ **then**
      $\Omega_K = \Omega_K - \{k^*\}$.
    end if
  end while
  for all n *in* $\mathcal{N}_{free}^{(\ell)}$ **do**
    for $i = (\ell + 1) : L$
      if $\lceil R_k^B / q \rceil \cdot q \ge x_{k,n}^{(\ell)} \cdot q \ \forall k$ **then**
        for $k = 1 : K$ **do**
          Set $x_{k,n}^{(i)} = x_{k,n}^{(\ell)}$, $1 \le k \le K$.
          Set $R_k^B = R_k^B - x_{k,n}^{(\ell)} \cdot q$.
          Set $P^{(i)} = P^{(i)} + \sum_k P_{k,n}^{(\ell)}$.
          Set $\widehat{R}_k = \max(\widehat{R}_k - x_{k,n}^{(\ell)} \cdot q, 0)$.
        end for
        let $\mathcal{N}_{free}^{(i)} = \mathcal{N}_{free}^{(i)} - \{n\}$.
      end if
    end for
  end for
end for ■

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their suggestions to improve the presentation of the paper.

REFERENCES

[1] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM system," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 171–178, Feb. 2003.

[2] V. K. N. Lau, "Optimal downlink space-time scheduling design with convex utility functions–multiple-antenna systems with orthogonal spatial multiplexing," *IEEE Trans. Veh. Technol.*, vol. 54, pp. 1322–1333, July 2005.

[3] H. Yin and H. Liu, "Performance of space-division multiple-access (SDMA) with scheduling," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 611–618, Oct. 2002.

[4] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-Based wireless broadband networks," *IEEE Commun. Mag.*, pp. 127–134, Dec. 2005.

[5] 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, pp. 1747–1758, Oct. 1999.

[6] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1150–1158, Nov. 2003.

[7] Y. J. Zhang and K. B. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1566–1575, Sept. 2004.

[8] S. Thoen, L. V. der Perre, M. Engels, and H. D. Man, "Adaptive loading for OFDM/SDMA-based wireless networks," *IEEE Trans. Commun.*, vol. 50, pp. 1798–1810, Nov. 2002.

[9] I. Koutsopoulos and L. Tassiulas, "Adaptive resource allocation in SDMA-based wireless broadband networks with OFDM signaling," in *Proc. IEEE Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002)*, vol. 3, June 2002, pp. 1376–1385.

[10] Y. M. Tsang and R. S. Cheng, "Optimal resource allocation in SDMA / muilti-input-single-output/OFDM systems under QoS and power constraints," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC-2004)*, vol. 3, Mar. 2004, pp. 1595–1600.

[11] D. Bartolome, A. I. Perez-Neira, and C. Ibars, "Practical bit loading schemes for multi-antenna multi-user wireless OFDM systems," in *Proc. Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, vol. 1, Nov. 2004, pp. 1030–1034.

[12] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2726–2737, Nov. 2005.

[13] 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, pp. 2726–2737, July 2005.

[14] W. Huang and K. B. Letaief, "A cross-layer resource allocation and scheduling for multiuser space-time block coded MIMO/OFDM systems," in *Proc. IEEE Internaltional Conference on Communication (ICC-2005)*, 2005, pp. 2655–2659.
[15] "Local and metropolitan area networks–part 16: air interface for fixed broadband wireless access systems," IEEE Standard Std. 802.16-2004.
[16] M. Shen, G. Li, and H. Liu, "Effective of traffic channel configuration on the orthogonal frequency division multiple access downlink performance," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1901–1913, July 2005.
[17] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, pp. 1218–1230, Oct. 1997.
[18] A. Conti, M. Z. Win, and M. Chiani, "Slow adaptive m-qam with diversity in fast fading and shadowing," *IEEE Trans. Commun.*, vol. 55, pp. 895–905, May 2007.
[19] ——, "Invertible bounds for m-qam in rayleigh fading," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1994–2000, Nov. 2005.
[20] K. G. Murty, *Operations Research.* Prentice Hall, 1995.
[21] H. Yaghoobi, "Scalable OFDMA physical layer in IEEE 802.16 WirelessMAN," *Intel Techonol. J.*, vol. 8, no. 3, 2004.
[22] 3GPP TR 25.892, "Feasibility study for OFDM for UTRAN enhancement," 3rd Generation Partnership Project, Tech. Rep., 2004-06.
[23] Universal Mobile Telecommunications System, "Selection procedures for the choice of radio transmission technologies of the UMTS," UMTS Std. 30.03, 1998.
[24] Z. Diao, D. Shen, and V. O. K. Li, "An adaptive packet scheduling algorithm in ofdm systems with smart antennas," in *Proc. Personal, Indoor and Mobile Radio Communications (PIMRC-2005)*, 2005, pp. 2151–2155.
[25] WiMAX forum, "Wimax system evaluation methodology," V.1.0, Tech. Rep., Jan. 2007.
[26] IEEE C802.16e-04/527r4, "Improved feedback for MIMO precoding," Tech. Rep.
[27] D. J. Love, R. W. Heath Jr., "What is the value of limited feedback for MIMO channels," *IEEE Commun. Mag.*, pp. 54–59, Oct. 2004.

Chun-Fan, Tsai was born in Miaoli, Taiwan. He received B.E. and M.E. degrees in department of communication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2004 and 2006, respectively. His research interests include radio resource management, scheduling, and voice over IP.

Chung-Ju Chang was born in Taiwan, ROC, in August 1950. He received the B.E. and M.E. degrees in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1972 and 1976, respectively, and the Ph. D degree in electrical engineering from National Taiwan University, Taiwan, in 1985. From 1976 to 1988, he was with 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. He was Director of the Institute of Communication Engineering from August 1993 to July 1995, Chairman of Department of Communication Engineering from August 1999 to July 2001, and Dean of the Research and Development Office from August 2002 to July 2004. Also, 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. Moreover, he serves as Editor for *IEEE Communications Magazine* and Associate Editor for *IEEE Transactions Vehicular Technology*. His research interests include performance evaluation, radio resources management for wireless communication networks, and traffic control for broadband networks. Dr. Chang is a member of the Chinese Institute of Engineers (CIE).

Fang-Chin Ren was born in Hsinchu, Taiwan. He received B.E., M.E., and Ph.D. degrees in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1992, 1994, and 2001, respectively. Since 2001, he has been a Protocol Design Engineer in Industrial Technology Research Institute, Taiwan, where he was involved in the design and development of WCDMA chipset, WiMAX mobile multihop relay technology, and 4G access technology. His current research interests include system performance analysis, protocol design, and mobile radio network.

Chih-Ming, Yen was born in Tainan, Taiwan. He is a Ph.D. student in department of communication engineering from National ChiaoTung University, Hsinchu, Taiwan. His research interests include radio resource management and wireless communication.