This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.

Adaptive Radio Resource Allocation for Downlink OFDMA/SDMA Systems

Chun-Fan Tsai*, Chung-Ju Chang*, Fang-Ching Ren†, and Chih-Ming Yen*
*Department of Communication Engineering
National Chiao Tung University
Hsinchu, Taiwan 300
Email: cjchang@cc.nctu.edu.tw
†Information and Communication Research Laboratories
Industrial Technology Research Institute
Hsinchu, Taiwan 300

***Abstract*—The paper proposes an adaptive radio resource allocation (ARRA) algorithm for downlink OFDMA/SDMA systems with goals to achieve quality of service (QoS) satisfied and throughput maximized. The ARRA algorithm considers multiple service classes of multimedia traffic and their diverse QoS requirements. It contains two parts, a dynamic priority adjustment scheme and a priority-based greedy (PBG) algorithm. The dynamic priority adjustment scheme gives high priority to urgent users and dynamically adjust the value frame by frame based on users' QoS requirement and queue occupancy. The PBG algorithm allocates the radio resource iteratively according to a cost value to maximize the system throughput while allocating enough resource to high-priority users. Simulation results show that the ARRA algorithm outperforms the conventional algorithms in terms of system throughput under the satisfaction of QoS requirements.**

I. Introduction

Orthogonal frequency division multiple access combined with space division multiple access (OFDMA/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. The SDMA with beamforming technique to multiplex multiple users on the same subchannel for increasing the system throughput. For a multiuser OFDMA system, the system data rate was maximized when each subcarrier was assigned to the user with the best channel gain [1]. However, when SDMA is enabled in an OFDMA system, the data rate of the system is maximized while the optimal set of cochannel users is selected for each subcarrier. When channel state information (CSI) is available at base station, a sophisticated radio resource allocation (RRA) scheme is needed for OFDMA/SDMA systems to maximize system throughput by exploiting system diversity.

In a modern wireless system that supports multimedia traffic, quality-of-service (QoS) guarantee should be a design consideration for RRA algorithms. An optimal resource allocation algorithm for OFDMA system was proposed to minimize the total transmission power consumption under the satisfaction of QoS requirement in [2]. The generalized processor sharing (GPS) scheduling was extended to the OFDM systems in [3]. The resource allocation for OFDMA/SDMA system was first investigated in [4] and [5]. However, in [6], the authors challenged the performance of [4] and [5] and proposed an optimal solution for maximizing information capacity. Also, practical bit loading schemes were proposed in [7], where a heuristic approach was taken to reduce the high complexity of the RRA algorithm in OFDMA/SDMA systems.

From these previous works, three observations can be induced. Firstly, all above schemes can be considered as fixed-priority schemes. The resource is either allocated to guarantee a fixed number of transmission bits or assigned according to predefined weights. Since the required resource is fixed in each OFDMA symbol, the time diversity is not well exploited and system throughput is degraded. Secondly, only bit error rate (BER) and/or minimum transmission rate were considered as QoS requirements in previous RRA algorithms. However, with the present of multimedia traffic, the delay requirement should also be included. Finally, most of researches assumed that a subcarrier is used as the basic allocation unit and each user always has data in its buffer. However, a subcarrier-based allocation is difficult to realize due to its high signaling overhead. Noticeably, the basic allocation unit in a practical OFDMA system (e.g. IEEE 802.16 [8]) is a subchannel, which is a set of subcarriers. Also, in realistic environments providing various service types, the traffic models should be taken into account in the design of RRA algorithms.

The paper proposes an adaptive radio resource allocation (ARRA) algorithm for downlink OFDMA/SDMA systems with an objective to maximize the system throughput and to guarantee the QoS of multimedia traffic. The ARRA algorithm is composed of two parts. The first part is an dynamic priority adjustment scheme, where priorities of users are dynamically adjusted frame by frame and the required resource of each user varies with time. By giving high priority to urgent users, it is believed that the ARRA algorithm can attain the tradeoff between throughput and QoS requirement better than the schemes with fixed priority. The second part of the ARRA algorithm is a low-complexity resource allocation scheme, called priority based greedy (PBG) algorithm, to allocate resource based on the priority obtained in the first

1-4244-0353-7/07/$25.00 ©2007 IEEE

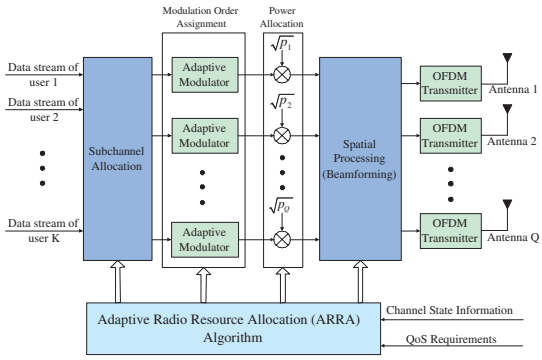

Fig. 1. Transmission structure of the OFDMA/SDMA System

part and the CSI of users. Simulation results show that the ARRA algorithm achieves the system throughput higher than conventional algorithms, under the QoS requirements.

II. System Model

The OFDMA/SDMA system is assumed to support three classes of service, real-time (RT), non-real-time (NRT), and best effort (BE), which are with different QoS requirements. For RT services, the QoS requirements consider BER, maximum delay tolerance, and maximum allowable 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 classes, and a traffic model is associated with the user. Denote these QoS requirements of BER, minimum required transmission rate, maximum delay tolerance, and maximum allowable dropping ratio by BER_k^*, R_k^*, D_k^*, and $P_{D,k}^*$ respectively. The system provides one individual queue for each downlink user at the base station. Packets of RT services will be dropped if the delay of packets exceeds the maximum delay tolerance, while packets of NRT services or BE services are allowed to be queued without being dropped if buffer occupancy is not overflowed.

The architecture of the downlink OFDMA/SDMA system with the ARRA algorithm is shown in Fig. 1, where data streams for K single-antenna mobile stations will be transmitted from the base station which is equipped with N subchannels and Q transmit antennas. A set of OFDM subcarriers forms an OFDMA subchannel, which is the basic unit for resource allocation and adaptive modulation in a practical OFDMA system [8]. A subchannel is assumed to have continuous b subcarriers since the grouping of continuous subcarriers results in highest multiuser diversity [9]. The time axis is divided into *frames* with fixed length, and each frame includes L OFDMA symbols for downlink transmission. Three possible modulation scheme QPSK, 16-QAM, and 64-QAM are using in this paper and let $q = 2 \times b$ be the number of transmission bits with the basic QPSK modulation over b subcarriers in one subchannel. The ARRA algorithm is executed at the beginning of every frame to properly allocate radio resource to all users according to their queue state, CSI, and QoS requirements. The radio resource allocation includes the subchannel allocation, modulation order assignment, power allocation, and beamforming control.

For the system under consideration, let Ψ_n be the set of the subcarriers in subchannel n and $\mathcal{K}_n^{(\ell)}$ be the set of users that are multiplexed on subchannel n for the ℓth OFDMA symbol. The transmit symbol vector in subcarrier i of subchannel n for the ℓth OFDMA symbol, denoted by $\mathbf{S}_i^{(\ell)}$, is given by

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

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 at subcarrier i at the ℓth OFDMA symbol. Notice that a normalized QAM modulation is used such that the data symbol has unitary mean energy.

A perfect CSI estimation for each user is assumed and the channel is fixed within a frame duration. Let $\mathbf{h}_{k,i}$ be a $1 \times Q$ vector denoting the frequency domain channel gain from base station to user k on subcarrier i. For simplicity and acceptable performance, the zero-force (ZF) transmit beamforming scheme [6], [7] is used. The cochannel users are orthogonal in space domain while ZF transmit beamforming is adopted. Therefore, 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 complex Guassian with zero mean and variance σ^2. From the above equation, the received SNR can be written as,

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

Note that the received SNR is affected by the beamforming vector. If the users with high spacial correlation are selected, the term $\mathbf{h}_{k,i} \mathbf{w}_{k,i}^{(\ell)}$, will be small and a poor received SNR will be resulted. A scheduler should select the users with low spacial correlation into the same subchannel.

If the cochannel user set, $\mathcal{K}_n^{(\ell)}$, is determined, the beamforming vectors of the users in subchannel n can be calculated by using the formulation of ZF beamforming. Then, it can be seen form (3) that the received SNR of user k depends on the allocated power, $\xi_{k,i}^{(\ell)}$. To maintain the BER performance, the allocated power to user k is set to the value such that the received SNR is equal to the minimum required SNR of user k, which can be obtained from 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 [10],

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

According, from (3) and (4), the allocated power, $\xi_{k,i}^{(\ell)}$, can be obtained by,

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

Thus the power allocated to user k on subchannel n for the ℓth OFDMA symbol, denoted by $p_{k,n}^{(\ell)}$, can be calculated as $p_{k,n}^{(\ell)} = \sum_{i\in\Psi_n} \xi_{k,i}^{(\ell)}$. In other words, the power allocated to a user should be sufficiently enough to guarantee the BER requirement if the user is selected by the ARRA algorithm.

III. Adaptive Radio Resource Allocation

Define $x_{k,n}^{(\ell)}$ as the assignment variable for indicating the transmission state of user k on subchannel n at the ℓth OFDMA symbol, where $x_{k,n}^{(\ell)} = 0, 1, 2, \text{or } 3$ means that no transmission, transmit using QPSK modulation, transmit using 16-QAM, and transmit using 64-QAM, respectively. Denote the assignment vector $\mathbf{x}^{(\ell)} \underset{T}{\equiv} \left[x_{1,1}^{(\ell)}, \cdots, x_{1,N}^{(\ell)}, \cdots, x_{k,1}^{(\ell)}, \cdots, x_{k,N}^{(\ell)}, \cdots, x_{K,1}^{(\ell)}, \cdots, x_{K,N}^{(\ell)}\right]$ the solution of the ARRA algorithm for the ℓth OFDMA symbol. Then, the allocated transmission bits to user k in this frame, denoted by R_k, and the cochannel sets, $\mathcal{K}_n^{(\ell)}$, can be obtained from the assignment vectors. Hence, if needed, $\mathcal{K}_n^{(\ell)}$ and $\mathbf{x}^{(\ell)}$ will be denoted by $R_k(\mathbf{x}^{(1)}\cdots\mathbf{x}^{(\ell)})$ and $\mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)})$ in the following. Also, $p_{k,n}^{(\ell)}$ can also be considered as a function of BER_k^* and $\mathbf{x}^{(\ell)}$.

Four constraints are considered on the design of the ARRA algorithm. The first one is the *subchannel allocation constraint*, which means that a subchannel can be allocated to Q users at most in order to make the ZF beamforming realizable. The constraint is expressed as $|\mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)})| \le Q$. The second one is the *total system power constraint*, which represents that the total power allocation for downlink data transmission have a limitation. Denote P_T the total power, the constraint is written as $\sum_{n=1}^{N}\sum_{k=1}^{K} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}) \le P_T$. For transmission efficiency, the third constraint, *buffer occupation constraint*, is added. The allocated bits to a user should not larger than the its buffer occupancy, which can be expressed as $R_k \le \lceil R_k^B/q \rceil \cdot q$, where R_k^B is the buffer length of user k.

For further satisfying the QoS requirement of the RT users and NRT users, the last constraint, *QoS fulfillment constraint*, is included. First, a priority value is set for each user at each frame according to its QoS requirements and queue occupancy. We define the priority value of user k, denoted by $\widehat{R}_k$, as the minimum number of transmission bits required at current frame in order to fulfill the user's QoS Requirements. Thus, the R_k should have the QoS fulfillment constraint expressed as $R_k \ge \widehat{R}_k$. At least $\widehat{R}_k$ bits should be allocated to user k at the current frame to guarantee its QoS requirements. Noticeably, the priority value of a user is dynamically adjusted frame by frame.

Therefore, the ARRA algorithm is formulated 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:

$$\begin{aligned} &|\mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)})| \le Q, \quad \forall n, \ell, \\ &\sum_{n=1}^{N}\sum_{k=1}^{K} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}) \le P_T, \quad \forall \ell, \\ &R_k \le \lceil R_k^B/q \rceil \cdot q, \quad \forall k, \\ &R_k \ge \widehat{R}_k, \quad \forall k. \end{aligned} \tag{6}$$

In this formulation, the system throughput is maximized under the four constraints. The optimization problem (6) can be easily solved by an integer programming method [11]. However, the complexity of the integer programming method grows exponentially with the number of users and is unacceptable in real applications. Hence, the proposed ARRA algorithm adopts a reduced-complexity approach based on greedy approach [11]. The proposed ARRA algorithm contains two parts to find the solution in (6), an dynamic priority adjustment scheme and a priority-based greedy (PBG) algorithm. The details are described in the following.

A. Dynamic Priority Adjustment Scheme

We here introduce a *time-to-expiration* (TTE) value, indicating the urgency degree of a user at the current frame. For user k, denote the TTE value and the number of residual bits of the head-of-line (HOL) packet by V_k and B_k, respectively. The smaller the V_k is, the more the degree of urgency of user k would be. For users with RT service class, the V_k is intuitively given by,

$$V_k = D_k^* - D_k, \tag{7}$$

where D_k is the time duration 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 in NRT service class, the V_k is given by,

$$V_k = \left\lfloor \frac{B_k' + B_k}{R_k^*} - D_k' \right\rfloor, \tag{8}$$

where D_k' is the the time duration while there is data buffered in the queue of user k before the current frame, B_k' is the total 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 (8) of NRT user k comes from the inequality $(B_k + B_k')/(V_k + D_k') \ge R_k^*$, which means that the average rate should be greater than the minimum required transmission rate. Finally, for users in BE service class, the V_k is intuitively set to be infinity.

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

$$\widehat{R}_k = \begin{cases} 0, & \text{if } V_k = \infty \\ \left\lceil \frac{B_k}{q} \right\rceil \cdot q, & \text{if } V_k \le V_{th} \\ \max\left(\left\lceil \frac{B_k}{V_k \cdot q} \right\rceil - \lceil \ln(V_k) \rceil, 0 \right) \cdot q, & \text{elsewise,} \end{cases} \tag{9}$$

where V_{th} is a threshold for V_k. If $V_k = \infty$, then it is intuitive to set $\widehat{R}_k$ as zero. If V_k is below the threshold V_{th}, it means that the degree of urgency of user k is very high in a fashion that the user k should complete its transmission in this current frame, i.e. $\widehat{R}_k = \left\lceil \frac{B_k}{q} \right\rceil \cdot q$. Otherwise, the design of $\widehat{R}_k$ is based on the average required transmit 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, so that the system can give the transmission opportunity to other high-urgent users. Note that a user with low priority could still be served by the base station if the channel quality of the user is good and other users with higher priority have been already served. Hence, the delay-tolerable users can take the advantage of time diversity by transmitting only when its channel is good. As for the threshold value V_{th}, it could be set to one if resource is always enough to satisfy $R_k \geq \widehat{R}_k$. However, since the user might be in cell boundary, the V_{th} could be set to a value greater than one to guarantee the QoS requirement earlier. In the later section of simulation, the V_{th} is set to three.

B. PBG Algorithm

The basic principle of the PBG algorithm is that every successive step is taken to minimize an immediate cost. The immediate cost is the increment of power of increasing one modulation order for a user on one subchannel. Denote $C_{k,n}^{(\ell)}$ the cost function of user k on subchannel n at the ℓth OFDMA symbol. If $0 \leq x_{k,n}^{(\ell)} \leq 2$, the cost is calculated as,

$$C_{k,n}^{(\ell)} = \sum_{k \in \mathcal{K}_n^{(\ell)}(\mathbf{x}^{+(\ell)})} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{+(\ell)}) - p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}), \tag{10}$$

where $\mathbf{x}^{+(\ell)}$ 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 $x_{k,n}^{+(\ell)} = x_{k,n}^{(\ell)} + 1$. Otherwise, $C_{k,n}^{(\ell)}$ is set to infinity since the maximum modulation order is reached and the modulation order can not be increased further. Increasing of modulation order from zero to one means adding a new user to a subchannel, which requires recalculation of beamforming vectors. The definition of the cost value also includes the increasing power for maintaining the same modulation order for the users that already in the subchannel. Hence the spatial correlation between the new user and the users that are already in the subchannel is also measured by the cost function.

The PBG algorithm is initialized as follows. The assignment variables are set to zero, and each user is given an instantaneous priority. Denote β_k the instantaneous priority of user k, and it is set as the priority from dynamic priority adjustment scheme. Denote $P^{(\ell)}$ the current used power and $\mathcal{N}_{free}^{(\ell)}$ the set of free subchannels for the ℓth OFDMA symbol. They are initialized as $P^{(\ell)} = 0$ and $\mathcal{N}_{free}^{(\ell)} = \{n | 1 \leq n \leq N\}$, $\forall \ell$. The PBG algorithm then sequentially allocates resource for each OFDM symbol used for downlink transmission in the current frame. In each symbol, two functions are performed, function ***Allocation-for-one-symbol*** and function ***Extend***. The PBG algorithm is depicted in the following pseudocode.

- ***PBG Algorithm***
  Set $\mathbf{x}^{(\ell)} = \mathbf{0}$, $\forall \ell$ *and* $\beta_k = \widehat{R}_k, \forall k$.
  Set $P^{(\ell)} = 0$ *and* $\mathcal{N}_{free}^{(\ell)} = \{n | 1 \leq n \leq N\}$, $\forall \ell$.
  for $\ell = 1 : L$ **do**
    Execute function ***Allocation-for-one-symbol***.
    Execute function ***Extend***.
  end for

In function ***Allocation-for-one-symbol***, an iterative algorithm is executed for resource allocation in symbol ℓ. A candidate user set, denoted by Ω, is constructed and an optimal pair of user and subchannel, (k^*, n^*), is selected, in every iteration. The Ω contains the backlogged users with highest instantaneous priority and the (k^*, n^*) is selected from the users in the candidate user set and the free subchannels such that the cost value, $C_{k^*,n^*}^{(\ell)}$, is minimum. If the power budget in the ℓth OFDMA symbol is still sufficient for increasing the modulation order for user k^* on subchannel n^*, then some states are updated as follows. The modulation order of the selected user on the selected subchannel is increased by one, i.e. $x_{k^*,n^*}^{(\ell)} = x_{k^*,n^*}^{(\ell)} + 1$. Additional q bits are allocated to the selected user in each iteration, and thus the queue length of user k^*, R_k^B, is decreased by q. Used power for the ℓth OFDMA symbol, $P^{(\ell)}$, is increased by the minimum cost. For fairness issue, the instantaneous priority of the user k^* is decreased by q until the priority become zero, i.e. $\beta_{k^*} = \max(\beta_{k^*} - q, 0)$. Thus, low-priority users can still have opportunity to be transmitted. The solution of resource allocation in the ℓth OFDMA symbol is given in vector $\mathbf{x}^{(\ell)}$ while this function is terminated.

In function ***Extend***, the same allocation could be extended over several OFDMA symbols to form a time burst transmission. This function is to reduce the signaling overhead of the system and the complexity of the PBG algorithm. If subchannel n the ℓth OFDMA symbol has been allocated to a specified group of users, according to the result of function ***Allocation-for-one-symbol***, then we can allocate the subchannel n in the $\ell+1$th OFDMA symbol to the same group of users. The same operation, which is called extension in this paper, can be done for symbol $\ell + 2$, $\ell + 3$, and so on. The extension would be performed as long as the current queue occupancy for each user in the specified group is not empty. The assignment variables, instantaneous priority, queue length, used power, and the set of free subchannels are updated if the extension is performed.

IV. Simulation Results and Discussion

The downlink OFDMA/SDMA system environments are configured according to the IEEE 802.16 standard [8], where parameters are listed in Table I. The path loss model is modeled as $128.1 + 37.6 \log R$ dB, where R is the distance between the base station and the user in kilometers [12]. The log-normal shadowing is assumed with zero mean and standard

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
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

TABLE II
THE QOS REQUIREMENT OF EACH TRAFFIC TYPE

	Required BER	Maximum Delay Tolerance	Maximum Allowable Dropping Ratio	Minimum Required Transmission Rate
Voice	10^{-3}	40ms	1%	- - -
Video	10^{-4}	10ms	1%	- - -
HTTP	10^{-6}	- - -	- - -	100 kbps
FTP	10^{-6}	- - -	- - -	- - -

deviation of 8 dB. The multipath channel for each antenna is modeled as six taps of Rayleigh-faded paths. Four kinds of traffic types are assumed in the system, voice traffic [13] of RT service, streaming video traffic of RT service [12], HTTP traffic of NRT service [12], and FTP traffic of BE service [12]. Note that channel coding is not used in the simulation for reducing simulation time. The number of users is increased from 40 to 600, and each traffic type has the same number of users. The traffic load is defined as the ratio of the total average arrival rate of all users over the system maximum transmission rate, which could be achieved when Q users are multiplexed for each subchannel and the highest modulation order is used. Table II lists the QoS requirements of each traffic type.

Three conventional RRA algorithms will be considered to compare with the proposed ARRA scheme. The first one is linkgain-based resource allocation (LBRA) [1], where resource is allocated to users according to users' CSI. For each subchannel, the first Q users with best channel quality are selected. The second one is multi-antenna multi-user maximum sum rate (MMSR) [7], which contains a user clustering procedure and a bit-removing algorithm. The former selects the set of cochannel users while the later determine the modulation orders of the select users. The last one is truncated generalized processor sharing (TGPS) [3], where resource is allocated to users based on predefined weights of all users. The predefined weight is set to 10, 5, and 1 for RT, NRT, and BE services, respectively.

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. The reasons are: the ARRA algorithm improves the system throughput by taking multiuser diversity and space domain correlation between users in its design. The system throughput of the MMSR scheme is near to

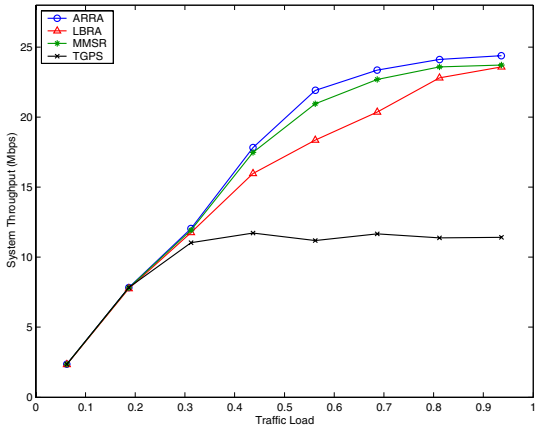

Fig. 2. System Throughput

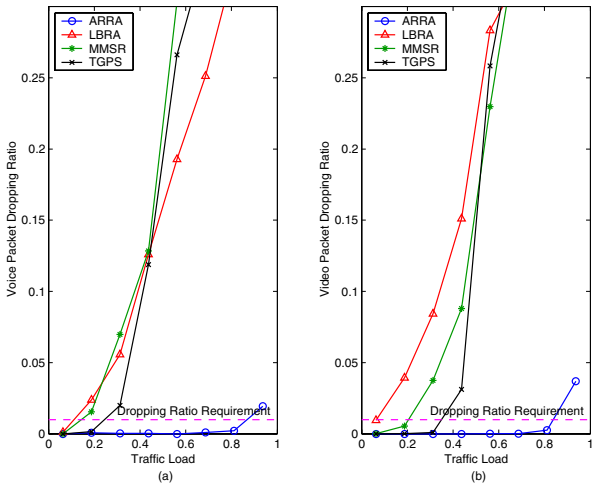

Fig. 3. (a) Packet Dropping Ratio of Voice Users (b) Packet Dropping Ratio of Video Users

that of the ARRA scheme because both of this two schemes takes throughput maximization as a design objective. The system throughput of LBRA scheme is less than that of the ARRA scheme for the reason that the optimal user grouping in space domain is not considered in LBRA algorithm. The system throughput of TGPS algorithm is smallest in the four algorithms; it is because the TGPS uses simplified algorithm for subchannel allocation and the multiuser diversity is not well exploited.

Figs. 3 (a) and 3 (b) depict the packet dropping ratio of voice users and the packet dropping ratio of video users, respectively. It can be seen that the voice/video packet dropping ratios of the ARRA algorithm are almost zero until the traffic load is greater than 0.8, while those of the other algorithms increase rapidly with the traffic load. The reason is that the LBRA or MMSR algorithm does not consider the QoS requirement of maximum delay tolerance for the RT users. As for the TGPS algorithm, since its maximum capacity is too small, its dropping ratio became large at high traffic load even though the TGPS algorithm gives large weights to RT users. On

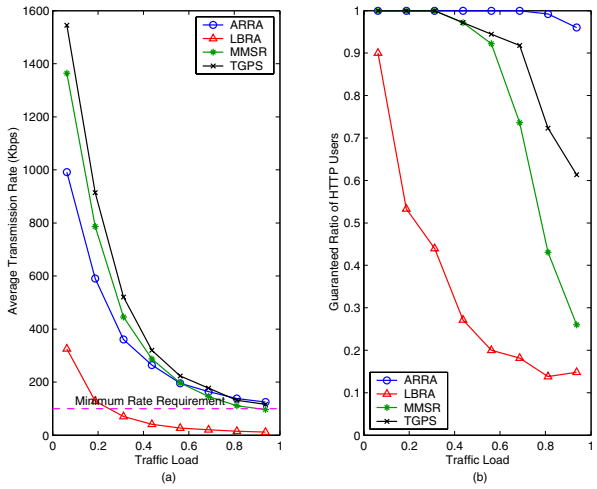

Fig. 4. (a) Average Transmission Rate of HTTP Users (b) Guaranteed Ratio of HTTP Users

the other hand, the ARRA algorithm considers the RT user with large packet delay as urgent users and gives the users high priority. Since resource is first provided for high-priority users, the delay requirement of the RT user is satisfied and the dropping ratio is the smallest.

Figs. 4 (a) and 4 (b) illustrate the average transmission rate of HTTP users and the guaranteed ratio of HTTP users, respectively. The guaranteed ratio of NRT users is defined as the ratio of the number of the QoS-satisfied HTTP users over total HTTP users. 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 transmission rate is guaranteed by giving high priority to the NRT users whenever their transmission rate is going to be lower than minimum required transmission rate. For the same reason, the guaranteed ratio of HTTP users is almost 100% in the ARRA algorithm when traffic load is low and still larger than 95% when the traffic load is 0.9. Although the average transmission rate of the MMSR or TGPS algorithm is higher than that in the ARRA algorithm, the guaranteed ratio drops earlier than the ARRA algorithm. For example, when traffic load is 0.8 the guaranteed ratio of the ARRA algorithm is 99% while that of TGPS algorithm and MMSR algorithm is only 70% and 40%, respectively. The LBRA algorithm has lowest guaranteed ratio of HTTP users since it only guarantees the transmission rate of the users with good channel quality.

V. Conclusions

In this paper, an adaptive radio resource allocation (ARRA) algorithm is proposed for downlink OFDMA/SDMA systems. The proposed ARRA algorithm contains a dynamic priority adjustment scheme to dynamically adjust the priority of users frame by frame, where NRT users with lower average transmission rate which is near the minimum required transmission rate and RT users which are with larger packet delay can be promoted to higher priority to obtain the enough resource earlier. It also consists of a PBG algorithm to efficiently allocate the resource based on a cost value. Simulation results show that the ARRA algorithm outperforms the conventional algorithms in term of system throughput, under the satisfaction of QoS requirements.

Acknowledgment

This work was supported by National Science Council, Taiwan, under contract number NSC 95-2752-E-009-014-PAE and Ministry of Education, Taiwan under Grants 95W803C.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] *Local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Standard Std. 802.16-2004.

[9] 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.

[10] 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.

[11] K. G. Murty, *Operations Research*. Prentice Hall, 1995.

[12] 3GPP TR 25.892, "Feasibility study for OFDM for UTRAN enhancement," 3rd Generation Partnership Project, Tech. Rep., 2004-06.

[13] Universal Mobile Telecommunications System, *Selection procedures for the choice of radio transmission technologies of the UMTS*, UMTS Std. 30.03, 1998.