OFDMA/SDMA 下鏈路系統之動態資源分配

Adaptive Radio Resource Allocation for Downlink OFDMA/SDMA Systems

研究生　：蔡俊帆　　Student : Chun-Fan Tsai

指導教授：張仲儒　博士　　Advisor: Dr. Chung-Ju Chang

國立交通大學

電信工程學系

碩士論文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in
Communication Engineering
June 2006
Hsinchu, Taiwan

中華民國九十五年六月

OFDMA/SDMA 下鏈路系統之動態資源分配

研究生：蔡俊帆　　指導教授：張仲儒

國立交通大學電信工程學系碩士班

摘 要

為了要達到高速無線傳輸，正交分頻多工(Orthogonal Frequency Division Multiplexing OFDM) 及多天線的架構是近年來在傳輸實體層(PHY)相當重要的研究領域。在這個系統之下，其多重擷取方式是一個 OFDMA/SDMA 的系統。另外，在現今多媒體通訊的環境下，傳輸服務品質(Quality of Service QoS)的保證是一個很重要的課題。因此，一個能夠有效使用系統資源同時能保證 QoS 的無線資源分配(Radio Resource Allocation) 演算法是必要的。

在本篇論文中，我們提出了一個動態無線資源分配(Adaptive Radio Resource Allocation ARRA)的演算法。其主要的目地是系統傳輸速率的最佳化及 QoS 的滿足。除此之外，多重服務層級的使用者(包含 real-time, non-real-time，及 best-effort)的多種 QoS Requirements，也是被考慮在內的。在提出的演算法中，我們給予較高緊迫性的使用者較高的 priority，並且動態調整每個使用者的 priority。我們也呈現了一個疊代式的演算法，將資源作有效率的分配。從模擬結果可以顯示出來我們所提出來的方法，不管是在系統傳輸速率的最佳化及 QoS 滿足方面，都比傳統的方法好。

Adaptive Radio Resource Allocation for Downlink OFDMA/SDMA systems

Student: Chun-Fan Tsai Advisor: Chung-Ju Chang

Department of Communication Engineering

National Chiao Tung University

Abstract

To support diverse quality of services (QoS) requirements for multimedia traffic, a well-designed radio resource allocation scheme is required to effectively utilize the system resource. The thesis proposes an adaptive radio resource allocation (ARRA) algorithm for downlink OFDMA/SDMA systems. The goals of ARRA algorithm are QoS satisfaction and throughput maximization. To serve the increasing multimedia traffic in today's network, diverse QoS requirements and multiple service classes, which include real-time (RT), non-real-time (NRT) and best effort (BE) services, are considered. In the proposed ARRA algorithm, we give high priority to the urgent users and dynamically adjust the priority value of users such that only the user really required resource is given a high priority value. Resource is first allocated to high priority users for QoS guarantee. An iterative algorithm, named priority based greedy (PBG) algorithm in this thesis, is presented to efficiently allocate the resource based on a cost value. From the simulation results, we can conclude that the ARRA algorithm outperforms the conventional algorithms in terms of system throughput and the satisfaction of QoS requirements.

誌　謝

能夠完成這篇碩士論文，我必須要感謝許多人。首先要感謝張仲儒教授的指導，沒有老師的在專業上引導以及待人處事方面的教誨，我不會有這兩年豐富的碩士生活。再來要感謝從工研院回來指導我們論文的芳慶學長，以及每當我遇困難時，給我寶貴建議的義昇學長，家慶學長，詠翰學長，立峰學長和文祥學長。感謝志明、凱元、朕逢、立忠及宗軒在我剛近實驗室的時候給我許多照顧。還有碩一的學弟妹們，建安、建興、佳璇、佳泓、正昕、世宏以及尚曄，希望你們未來能寫出好的論文。當然，一定要感謝和我一起走過兩年的家源、煖玉和琴雅，感謝你們兩年來的陪伴，很高興我們一起畢業了。

最後，我要感謝我的父母和家人，沒有他們在背後的支持和照顧，我不會有完成碩士學位的一天。

蔡俊帆　謹誌

民國九十五年七月

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Orthogonal frequency division multiple access (OFDMA) combined with space division multiple access (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 beamforming technique in an multiple-antenna system and multiplexes multiple users on the same subchannel to increase the system throughput.

For a multiuser OFDMA system, Jang and Lee proved that the data rate was maximized when each subcarrier was assigned to the user with the best channel gain [1]. However, the above statement is not always true when SDMA is enabled in an OFDMA system. Instead, the data rate of the system is maximized while the optimal set of cochannel users is selected for each subcarrier. The selection of cochannel users depends on the spacial signature of each user, and the algorithm for finding the optimal set of cochannel users requires high computational complexity [2, 3]. Hence, when channel state information (CSI) is available at 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 system throughput is maximized.

On the other hand, in a modern wireless system that supports multimedia traffic, quality-of-service (QoS) guarantee should be a design consideration for RRA algorithms. Wong et

al. [4] proposed an algorithm for OFDMA system to minimize the total transmission power consumption under the satisfaction of QoS requirement, which was defined as a specified data transmission rate and bit error rate (BER). Computational efficient methods for reducing the complexity in [4] are presented in [5, 6]. Thoen et al. [7] investigated the same optimization problem as [4] but with the presence of SDMA. However, the 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 [8] first considered the rate maximization problem, where they tried to maximum the SIR (Signal-to-Interference Ratio) for the cochannel users, without the constraint of QoS requirement. For the constrained case, they also proposed an algorithm for rate maximization under the same QoS requirement as in [4]. However, the power is fixed in each subcarrier and the SDMA is not enabled. In [9], Tsang and Cheng challenged the performance of [7] and [8] and proposed an optimal solution for maximizing information capacity. Also, practical bit loading schemes were proposed in [10], where a heuristic approach was taken to reduce the high complexity of the RRA algorithm in OFDMA/SDMA systems.

Two kinds of optimization problem formulation for resource allocation in OFDM-based system can be found in the literature, namely (i) power minimization [4,7] and (ii) rate maximization [9,10]. The former tried to minimum the overall transmit power given constraints on users' data rate or BER, while the latter tried to maximum the system data rate with constraints on the total power budget and users' BER performance. However, the fairness issue in the above two problem formulations was ignored. A new formulation to maximize the data rate with the constraint of that the allocated rate for each user is proportional to a specified weight was proposed [11]. Cai, Shen, and Mark [12], studied a resource management scheme for packet transmission in OFDM wireless communication systems, and they extended the generalized processor sharing (GPS) scheduling to OFDM systems. 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.

Based on the previous works, three observations can be obtained. Firstly, 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 allocated 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, which results in throughput degradation. As show in [6], the system throughput increases with the number of users due to multiuser diversity but decreases while the number of users is further increased. 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. Most packets should be transmitted within its delay bound, otherwise the packet is dropped. A RRA algorithm should make the dropping rate below an desired level. And thirdly, most of researches used theoretical assumption, where a subcarrier can be used as the basic allocation unit and each user always has data in its buffer. However, a subcarrier-based allocation are difficult to realize due to its high signaling overhead. Noticeably, the basic allocation unit in a practical OFDMA system (e.g. IEEE 802.16 [13]) is a subchannel, which is a set of subcarriers. Moreover, in more 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. In this paper, the priority is dynamically adjusted with time so that the required resource of each user varies with time. By giving high priority to urgent users, it is believed that the tradeoff between throughput and QoS requirement would be

better than the schemes with fixed priority. The ARRA algorithm is to maximize the system throughput and to satisfy user's QoS requirements. It is composed of two parts, the first part is an dynamic priority adjustment scheme, where a priority is given to each user based on its QoS requirements and its present queue state. And the second part is a low-complexity resource allocation scheme, based on the priority obtained in first part and CSI of all users, to maximize system throughput under the total power constraint.

The thesis is organized as follows. The system model of the considered OFDMA/SDMA system is introduced in chapter 2. In chapter 3, the details of the proposed algorithm is presented. Chapter 4 discusses the performance of proposed algorithm through simulation. Finally, conclusions are given in chapter 5.

Chapter 2

System Model

2.1 Services

The OFDMA/SDMA system is assumed to support three service classes, real-time (RT) services, non-real-time (NRT) services, and best effort (BE) services, which have different QoS requirements. For RT services, BER, maximum delay tolerance, and maximum allowable dropping ratio are considered as QoS requirements. For NRT services, BER and minimum required transmission rate are considered. For BE services, only BER is included in the QoS requirement. Each user belong to one kind of service classes, and a corresponding set of QoS requirements and a traffic model are associated with the user. Denote these QoS requirements of BER, minimum required transmission rate, maximum delay tolerance, and maximum allowable dropping rate by BER_k^*, R_k^*, D_k^*, and $P_{D,k}^*$ respectively.

Four kinds of traffic types are assumed in the system. The first one is voice traffic [14] of RT service. Each voice user traffic is modeled as an ON-OFF model, in which the length of ON period and OFF period follows the exponential distribution. Packets are generated during ON period with constant rate and none during OFF period. The second traffic type is the streaming video traffic [15] of RT service. Each frame of video data arrives at a regular

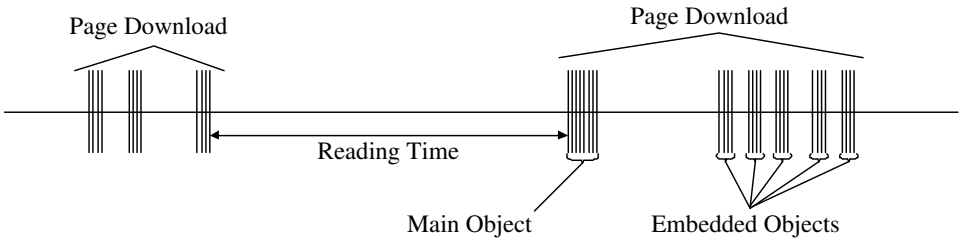


Figure 2.1: Packet Trace in Typical HTTP traffic

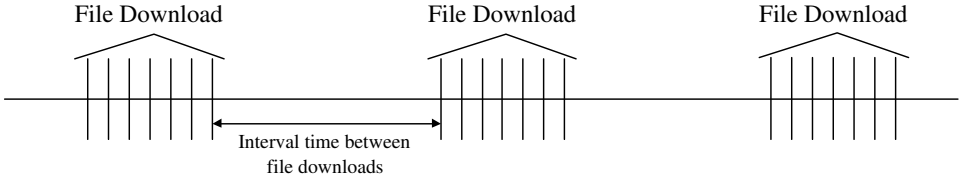
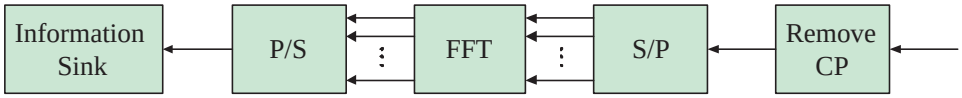


Figure 2.2: Packet Trace in a Typical FTP traffic

interval determined by the number of frames per second. Each frame is decomposed into a fixed number of slices, each transmitted as a single packet. The size of these packets is distributed in a truncated Pareto distribution. Encoding delay at the video encoder introduces delay intervals between the packet of a frame. These intervals are modeled by a truncated Pareto distribution.

The third one is HTTP traffic type [15] of NRT service, where the behavior of web browsing is modeled. Fig 2.1 shows the packet trace of typical HTTP traffic, where the traffic of HTTP user is modeled as a sequence of page downloads. Each page is composed of a main object and several embedded objects and each object is divided into several packets. The interval between download pages, which represents reading time in web browsing, is distributed in exponential distribution. The last one is the FTP traffic type [15] of BE service. Fig 2.2 shows the packet trace of typical FTP traffic, where each FTP user data is modeled as a sequence of file downloads. The size of a file is distributed as a truncated lognormal distribution, and the interval time between files is distributed in exponential distribution.

(a) OFDM Transmitter



(b) OFDM Receiver

Figure 2.3: OFDM Transceiver for Single User

The base station provides one individual queue for each user. Arriving packets from each user are stored in its own queue in a first-in first-out manner. The packets of RT services will be dropped if the delay of the packets exceed the maximum delay tolerance. On the other hand, the packets of NRT services or BE services are allowed to be queued without being dropped if buffer occupancy is not overflowed. Also, retransmission due to erroneous transmission of packets is not considered in this paper.

2.2 OFDM Transceiver

Fig. 2.3(a) and 2.3(b) shows the OFDM transmitter block diagram and OFDM receiver block diagram, respectively. An OFDM transmitter separates serial input information symbols into a parallel form and fed the parallel information symbols into corresponding subcarriers. In the discrete time domain, an OFDM modulator can be easily implemented by inverse fast Fourier transform (IFFT) and a parallel-to-serial (P/S) converter. In order to cancel the inter-carrier interference (ICI) due to multipath, a cyclic prefix (CP) is attached, which is a copy of the last portion of OFDM signal, to the front of itself.

In contrast to OFDM transmitter, an OFDM receiver removes the CP first, and then converts serial receive symbols into parallel form for performing the fast Fourier transform (FFT) as a demodulator. Finally, the output of FFT block will be fed into a P/S converter.

If the length of CP is lager than the delay spread of the channel between transmitter and receiver, each receive symbol is equal to the transmit symbol multiplied by an frequency channel gain. In this case, the OFDM system can effective mitigate the effect of multipath in typical wireless channel.

2.3 OFDMA/SDMA System

In the multiuser OFDMA/SDMA environment, the system is a space-time-frequency multiplexing system. Users are multiplexed in time domain and frequency by transmitting at different OFDMA symbols and different subcarriers, respectively. In space domain, users' data are separated by using beamforming scheme. A set of OFDM subcarriers are grouped into an OFDMA subchannel as a basic allocation unit in frequency domain. Theoretically, the mapping form subcarriers into subchannel can be arbitrary. In practice, however, regular mapping is often used for easy implementation [13,16]. Also, it has been shown that grouping of continuous subcarriers would result in highest multiuser diversity [16], which maximizes the system throughput. Hence, a subchannel is assumed to have continuous b subcarriers. One subchannel in frequency domain by one OFDMA symbol in time domain forms the basic unit for resource allocation and adaptive modulation. Since it is a multiple-antenna system, this time-frequency block can be allocated to more than one users and the users are multiplexed in space domain.

The architecture of the downlink OFDMA/SDMA system with the ARRA algorithm is shown in Fig. 2.4, where data streams for K single-antenna mobile stations (MS) will be transmitted from a base station (BS) which is equipped with N subchannels and Q transmit antennas. The time axis is divided into *frames* with fixed length, and each frame include L OFDMA symbols for downlink transmission. The ARRA algorithm is executed every frame time interval to properly allocate radio resource to all users according to their queue state,

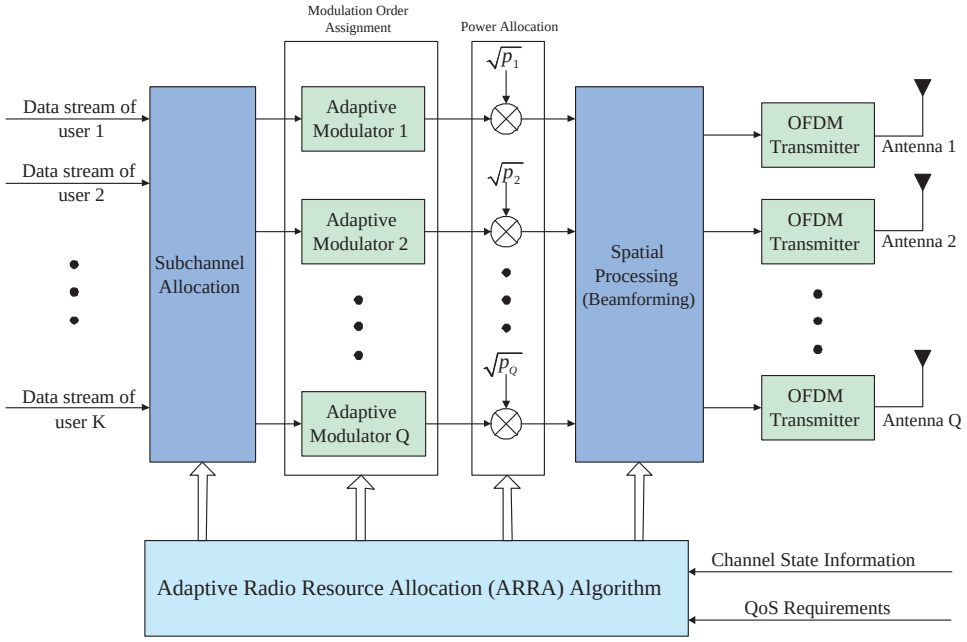


Figure 2.4: Transmission Structure of The OFDMA/SDMA Systems

CSI, and QoS requirements. The ARRA algorithm performs the subchannel allocation, modulation order assignment, power allocation, and beamforming control. Note that the results of subchannel allocation and modulation order assignment are specified in a frame header.

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 the users that are multiplexed on subchannel n. The transmit symbol vector in subcarrier i of subchannel n for the ℓth OFDMA symbol, denoted by $\mathbf{S}_i^{(\ell)}$, is given by

$$\mathbf{S}_i^{(\ell)} = \sum_{k \in \mathcal{K}_n^{(\ell)}} \sqrt{\xi_{k,i}^{(\ell)}} d_{k,i}^{(\ell)} \mathbf{w}_{k,i}^{(\ell)}, \qquad i \in \Psi_n, \tag{2.1}$$

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. Notice 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, and hence the channel is considered as fixed in a frame duration. The channel information can be obtained by using the reciprocal assumption in time division duplex (TDD) mode or by the feedback information in frequency division duplex (FDD) mode. Thus, a perfect CSI estimation for each user is assumed in this paper. Let $\mathbf{h}_{k,i}$ be a $1 \times Q$ vector denoting the frequency domain channel gain from BS to user k on subcarrier i. Note that $\mathbf{h}_{k,i}$ is not function of ℓ since the channel is fixed within a frame. The received signal of user k in subcarrier i for the ℓth OFDMA symbol, denoted by $Y_{k,i}^{(\ell)}$, is given by,

$$Y_{k,i}^{(\ell)} = \mathbf{h}_{k,i}\mathbf{S}_i^{(\ell)} + Z_{k,i}^{(\ell)} = \mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(\ell)}\sqrt{\xi_{k,i}^{(\ell)}}d_{k,i}^{(\ell)} + \sum_{u\in\mathcal{K}_n^{(\ell)},u\neq k} \mathbf{h}_{k,i}\mathbf{w}_{u,i}^{(\ell)}\sqrt{\xi_{u,i}^{(\ell)}}d_{u,i}^{(\ell)} + Z_{k,i}^{(\ell)}, \tag{2.2}$$

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. The second term in the right hand side of (2.2) is the *multi-beam interference* due to simultaneous transmission of independent data streams. From (2.2), the received signal-to-interference-ratio (SINR) of user k in subcarrier i for the ℓth OFDMA symbol, denoted by $SINR_{k,i}^{(\ell)}$, can be obtained by,

$$SINR_{k,i}^{(\ell)} = \frac{\xi_{k,i}^{(\ell)}\left|\mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(\ell)}\right|^2}{\sum_{u\in\mathcal{K}_n^{(\ell)},u\neq k}\xi_{u,i}^{(\ell)}\left|\mathbf{h}_{u,i}\mathbf{w}_{u,i}^{(\ell)}\right|^2 + \sigma^2}. \tag{2.3}$$

Hence, the receive SINR is a function of channel gain, user grouping, and beamforming vectors.

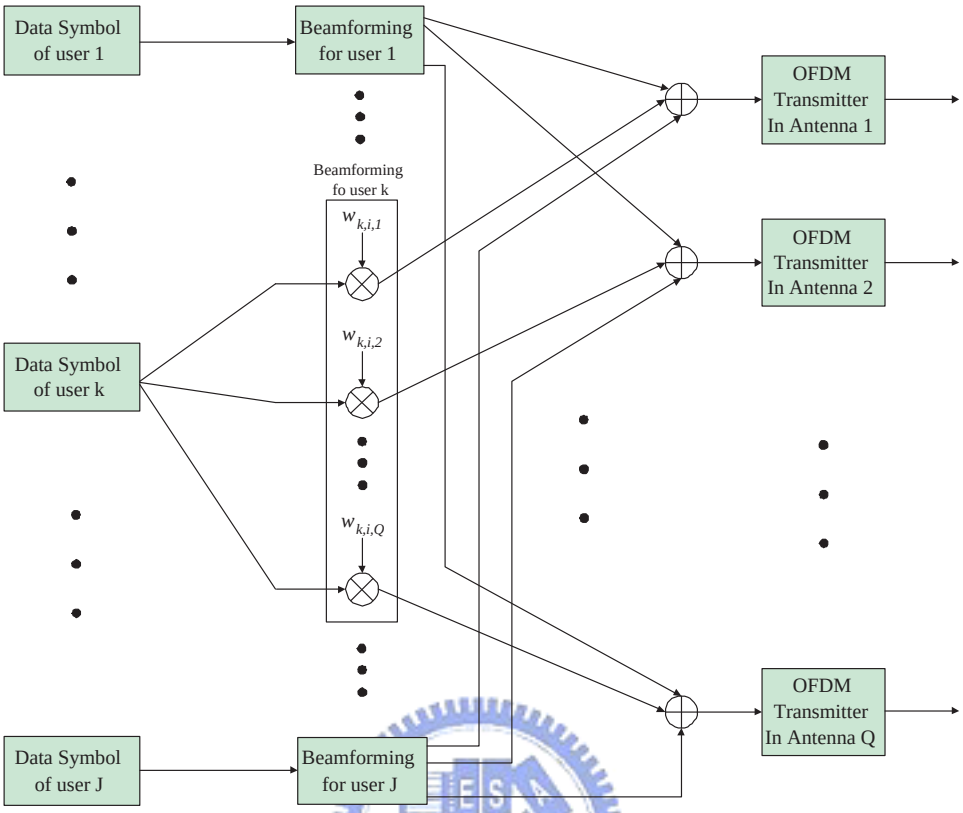


Figure 2.5: Example of Transmit Beamforming

2.4 Transmit Beamforming for SDMA

Fig 2.5 shows an example of transmit beamforming in subcarrier i of subchannel n, where J users multiplexed on subchannel n and $\mathcal{K}_n^{(\ell)} = \{1, \cdots, k, \cdots J\}$. The data symbol of a user is multiplied by different weights for transmitting at different antennas. The incoming signal of different users for the same antenna is sum together and fed into the OFDM transmitter in that antenna. Note that the design of beamforming vectors influence the interference between the users and the performance of the system. In this thesis, the zero-force (ZF) transmit beamforming scheme [2,9,10] is used for its simplicity and acceptable performance. Also, the performance of ZF beamforming is equivalent to the minimum mean square error (MMSE) beamforming for a low number of cochannel users or high SNR [7].

For a given $\mathcal{K}_n^{(\ell)}$, the formulation of ZF beamforming vectors, $\mathbf{w}_{u,i}^{(\ell)}$, can be expressed as

$$\begin{aligned} &\mathbf{h}_{k,i}\mathbf{w}_{u,i}^{(\ell)} = 0, \quad \forall u, k \in \mathcal{K}_n^{(\ell)}, u \neq k, \qquad i \in \Psi_n, \\ &\text{subject to } ||\mathbf{w}_{u,i}^{(\ell)}|| = 1, \forall u \in \mathcal{K}_n^{(\ell)}. \end{aligned} \tag{2.4}$$

The solution of (2.4) can be easily obtained by using psedoinverse and a normalization factor. Let $\mathbf{W}_i^{(\ell)} = [\mathbf{w}_{1,i}^{(\ell)}, \cdots, \mathbf{w}_{k,i}^{(\ell)}, \cdots, \mathbf{w}_{J,i}^{(\ell)}]$ be an $Q \times J$ beamforming matrix in subcarrier i of subchannel n, $i \in \Psi_n$. Denote $\mathbf{H}_i(\mathcal{K}_n^{(\ell)})$ the $J \times Q$ channel matrix of the users in $\mathcal{K}_n^{(\ell)}$, and thus,

$$\mathbf{H}_i(\mathcal{K}_n^{(\ell)}) = [\mathbf{h}_{1,i}, \cdots, \mathbf{h}_{k,i}, \cdots, \mathbf{h}_{J,i}]^T. \tag{2.5}$$

Then, the ZF beamforming matrix of subcarrier i for the ℓth OFDMA symbol, $\mathbf{W}_i^{(\ell)}$, is given by,

$$\mathbf{W}_i^{(\ell)} = \mathbf{H}_i^{\dagger}(\mathcal{K}_n^{(\ell)}) \left(\mathbf{H}_i(\mathcal{K}_n^{(\ell)})\mathbf{H}_i^{\dagger}(\mathcal{K}_n^{(\ell)})\right)^{-1} \Lambda_i, \tag{2.6}$$

where $\mathbf{H}_i^{\dagger}(\mathcal{K}_n^{(\ell)})$ denote the complex conjugate transpose of $\mathbf{H}_i(\mathcal{K}_n^{(\ell)})$, and Λ_i is a diagonal matrix for making the beamforming matrix unitary.

With the ZF transmit beamforming, the $Y_{k,i}^{(\ell)}$ in (2.2) will turn out to be

$$Y_{k,i}^{(\ell)} = \mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(\ell)}\sqrt{\xi_{k,i}^{(\ell)}}d_{k,i}^{(\ell)} + Z_{k,i}^{(\ell)}, \tag{2.7}$$

and $SINR_{k,i}^{(\ell)}$ in (2.3) becomes

$$SINR_{k,i}^{(\ell)} = \frac{\xi_{k,i}^{(\ell)} \left| \mathbf{h}_{k,i} \mathbf{w}_{k,i}^{(\ell)} \right|^2}{\sigma^2}. \tag{2.8}$$

From the above derivation, we can see that different user grouping, i.e. different $\mathcal{K}_n^{(\ell)}$, makes different channel matrix, $\mathbf{H}_i(\mathcal{K}_n^{(\ell)})$, and hence different beamforming matrix, $\mathbf{W}_i^{(\ell)}$. And from (2.8), $SINR_{k,i}^{(\ell)}$ is a function of $\mathbf{w}_{k,i}^{(\ell)}$. Hence, the received SINR value depended on the selection of cochannel users, i.e. the construction of $\mathcal{K}_n^{(\ell)}$. If the users with high spacial correlation are selected, the effective linkgain, which is defined as the term $\mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(\ell)}$, will be small and a poor received SINR will be resulted. A scheduler should select those cochannel users such that their effective linkgains are large enough.

2.5 Power Allocation

The allocated power to user k on subcarrier i for the ℓth OFDMA symbol, $\xi_{k,i}^{(\ell)}$, is determined by the minimum required SINR of user k, which can be obtained from BER_k^* and the modulation scheme of user k. A tight bound of the BER of when using M-QAM modulation is given by [17]

$$BER \leq 0.2e^{-1.5\frac{SINR}{M-1}}. \tag{2.9}$$

Thus, if user k adopts M-QAM modulation, the minimum required SINR, $SINR_k^*$, is given by,

$$SINR_k^* = -\frac{\ln(5\, BER_k^*)}{1.5}(M-1). \tag{2.10}$$

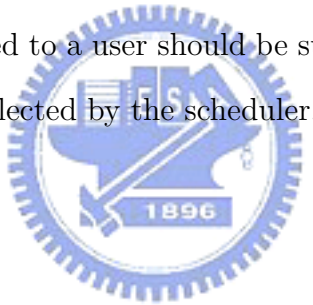
The allocated power, $\xi_{k,i}^{(\ell)}$, is set to the value such that the received SINR in (2.8) is equal to the required SINR in (2.10), hence it can be obtained by,

$$\xi_{k,i}^{(\ell)} = \frac{-\ln(5\,BER_k^*)(M-1)}{1.5} \frac{\sigma^2}{\left|\mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(\ell)}\right|^2}. \tag{2.11}$$

Thus the allocated power to user k on subchannel n for the ℓth OFDMA symbol, denoted by $p_{k,n}^{(\ell)}$, can be obtained by,

$$p_{k,n}^{(\ell)} = \sum_{i \in \Psi_n} \xi_{k,i}^{(\ell)}. \tag{2.12}$$

In other words, the power allocated to a user should be sufficiently enough to guarantee the BER performance if the user is selected by the scheduler.



Chapter 3

Resource Allocation Algorithms for OFDMA/SDMA Systems

Define $x_{k,n}^{(\ell)}$ as the assignment variable which indicates the modulation order of user k on subchannel n for the ℓth OFDMA symbol. where $x_{k,n}^{(\ell)} \in \{0, 1, 2, 3\}$, $1 \leq k \leq K$, $1 \leq n \leq N$, and $1 \leq \ell \leq L$. The state of transmission for user k at subchannel n in the ℓth OFDMA symbol is determined by the assignment variable as show in the following table.

Value	Meaning
$x_{k,n}^{(\ell)} = 0$	No transmission
$x_{k,n}^{(\ell)} = 1$	Transmit using QPSK modulation
$x_{k,n}^{(\ell)} = 2$	Transmit using 16-QAM modulation
$x_{k,n}^{(\ell)} = 3$	Transmit using 64-QAM modulation

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. Then, the allocated bits

to user k in this frame, denoted by R_k, can be calculated from $\mathbf{x}^{(\ell)}$, $1 \leq \ell \leq L$, by

$$R_k = R_k(\mathbf{x}^{(1)} \cdots \mathbf{x}^{(\ell)} \cdots \mathbf{x}^{(L)}) = \sum_{\ell=1}^{L} \sum_{n=1}^{N} q \cdot x_{k,n}^{(\ell)}, \tag{3.1}$$

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 for the ℓth OFDMA symbol, $\mathcal{K}_n^{(\ell)}$, can be obtained from $\mathbf{x}^{(\ell)}$ by,

$$\mathcal{K}_n^{(\ell)} = \mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)}) = \left\{ k | x_{k,n}^{(\ell)} > 0, 1 \leq k \leq K \right\}. \tag{3.2}$$

From (2.11) and (2.12), 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.

3.1 Adaptive Ration Resource Allocation

3.1.1 Design Constraints and Problem Formulation

The adaptive radio resource allocation (ARRA) algorithm is to determine an optimal assignment vector for each OFDMA symbol in a frame such that the total system throughput is maximized while each user's QoS requirements is satisfied. On the design of the ARRA algorithm for the downlink OFDMA/SDMA systems, four constraints are considered.

1. *Subchannel Allocation Constraint:* A subchannel can be allocated to Q users at most for each OFDMA symbol in the OFDMA/SDMA systems. If more than Q users are multiplexed on a subchannel, then the ZF transmit beamforming would be unrealized.

The constraint is expressed as,

$$|\mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)})| \leq Q, \quad \forall n, \ell. \tag{3.3}$$

Also, three possible modulation orders, QPSK, 16-QAM, and 64-QAM, may be assigned to the scheduled users.

2. *Total System Power Constraint:* The power allocation for downlink data transmission should have a limitation at the transmitter of BS. Denote P_T the total system power constraint for every OFDMA symbol, and the total power allocation is bounded as,

$$\sum_{n=1}^{N}\sum_{k=1}^{K} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}) \leq P_T, \quad \forall \ell. \tag{3.4}$$

3. *Buffer Allocation Constraint:* For transmission efficiency, the allocated transmission bits to user k in a frame, R_k, should not larger than the user's buffer occupancy at the beginning of this frame, denoted by R_k^B (bits). Since the basic allocation unit is a subchannel, the upper bound of R_k should be precisely expressed by,

$$R_k \leq \left\lceil R_k^B/q \right\rceil \cdot q, \quad \forall k. \tag{3.5}$$

4. *QoS Fulfillment Constraint:* For further satisfying the QoS requirement of maximum delay for RT users and the QoS requirement of minimum required transmission rate for NRT users, a priority value is set for each user according to its QoS requirements and queue state. We define the priority value of user k, denoted by $\widehat{R}_k$, as the minimum number of transmission bits required at current frame otherwise the user's QoS

requirements cannot be fulfilled. The larger the $\widehat{R}_k$ is, the higher the priority is and the more the resource should be provided to user k. We set the QoS fulfillment constraint expressed as,

$$R_k \geq \widehat{R}_k, \quad \forall k. \tag{3.6}$$

This means that at least $\widehat{R}_k$ bits should be allocated to user k at the current frame in order to satisfy its QoS requirements. Noticeably, the priority value of user is dynamically adjusted frame by frame.

Therefore, the ARRA algorithm is formulated as an optimization problem given by,

$$\begin{aligned}
(\mathbf{x}^{*(1)} \cdots \mathbf{x}^{*(L)}) = \arg \max_{\mathbf{x}^{(1)} \cdots \mathbf{x}^{(L)}} & \sum_{k=1}^{K} R_k(\mathbf{x}^{(1)} \cdots \mathbf{x}^{(L)}) \\
\text{subject to the following constraints:} & \\
& |\mathcal{K}_n^{(\ell)}(\mathbf{x}^{(\ell)})| \leq Q, \quad \forall n, \ell, \\
& \sum_{n=1}^{N} \sum_{k=1}^{K} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}) \leq P_T, \quad \forall \ell, \\
& R_k \leq \left\lceil R_k^B / q \right\rceil \cdot q, \quad \forall k, \\
& R_k \geq \widehat{R}_k, \quad \forall k.
\end{aligned} \tag{3.7}$$

In this formulation to find the optimal set, $\left\{\mathbf{x}^{*(\ell)} | 1 \leq \ell \leq L\right\}$, the system throughput is maximized under the satisfaction of system constraints. The optimization problem (3.7) can be easily solved by an integer programming method [18]. However, the complexity of the integer programming method grows exponentially with the number of users. Such a complexity is infeasible for real-time implementation when there are many users in the system. Hence, a reduced-complexity approach based on greedy algorithm [5, 18] is adopted

for the proposed ARRA algorithm. The greedy approach has been shown that it can achieve a near-optimal solution with lower computational complexity. Since the priority value of user is changed frame by frame, the proposed ARRA algorithm contains two parts to find the solution in (3.7). The first part is an dynamic priority adjustment scheme and the second part is a priority-based greedy (PBG) algorithm. The former gives each user a suitable priority value while the latter allocates radio resource to users with reasonable complexity. The details are described in the following.

3.1.2 Dynamic Priority Adjustment Scheme

Owing to the traffic burst characteristic and diverse QoS requirements, users' required resource varies with time. Accordingly, the ARRA algorithm dynamically adjusts the priority value of each user frame by frame, and the allocation of radio resource to users is according to their priority. We here introduce a *time-to-expiration* (TTE) value, indicating the degree of urgency of a user at the current frame. Denote the TTE value and the number of residual bits of the head-of-line (HOL) packet of user k 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 given by,

$$V_k = D_k^* - D_k, \tag{3.8}$$

where D_k is the time duration from the arrival of the HOL packet of user k to the present frame, and the unit of both D_k and D_k^* is in frames. The derivation of V_k of RT user k is directly from the delay requirement of RT users, which requires that the total delay should be smaller than required delay, i.e. $D_k + V_k \leq D_k^*$. Thus, a RT user's HOL packet should complete its transmission within its TTE value. Otherwise, the delay requirement of the

user is not satisfied, and the packet is dropped.

For users in NRT service class, the V_k is given by,

$$V_k = \left\lfloor \frac{B'_k + B_k}{R^*_k} - D'_k \right\rfloor, \tag{3.9}$$

where $\lfloor x \rfloor$ is the largest integer smaller than x, D'_k is the past active period of user k in unit of frames, B'_k is the total transmission bits of user k in its past active period, and R^*_k is the minimum required transmission rate in a unit of bits per frame. The past active period of a user is defined as the time duration when there is data buffered in the user's queue before the current frame. The derivation of V_k in (3.9) 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. Hence, similar to RT users, a 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 in BE service class, the V_k is set to be infinity since there is no delay or rate requirement for BE users.

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

$$\widehat{R}_k = \begin{cases} 0, & \text{if } V_k = \infty \\ \left\lceil \frac{B_k}{q} \right\rceil \cdot q, & \text{if } V_k \leq V_{th} \\ \max\left(\left\lceil \frac{B_k}{V_k \cdot q} \right\rceil - \lceil \ln(V_k) \rceil, 0 \right) \cdot q, & \text{elsewise,} \end{cases} \tag{3.10}$$

where $\lceil x \rceil$ is the smallest integer larger than x and 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 such a fashion that the user k should complete its transmission in this current frame, thus $\widehat{R}_k$ is set to equal to $\left\lceil \frac{B_k}{q} \right\rceil \cdot q$. Otherwise, the derivation 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, which are with large V_k, in order to 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 users with higher priority are already severed and the channel quality of the user is good. Hence, the delay-tolerable users can take the advantage of time diversity by transmitting only when its channel is good, and this can increase the system throughput. The threshold value V_{th} 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.

3.1.3 PBG Algorithm

The basic principle of the PBG algorithm is that every successive step is taken to minimize an immediate cost. We define the immediate cost as the increment of power of increasing one modulation order for a user on one subchannel. In each iteration of the PBG algorithm, an optimal pair of user and subchannel is selected and the modulation order of the user on that subchannel is tried to increase. The optimal pair is defined as the high-priority user on the subchannel such that its cost value is smallest. To satisfy QoS requirements, only users with high priority can be selected. The cost function of user k on subchannel n at the ℓth OFDMA symbol, denoted by $C_{k,n}^{(\ell)}$, is defined as,

$$C_{k,n}^{(\ell)} = \begin{cases} \sum_{k \in \mathcal{K}_n^{(\ell)}(\mathbf{x}^{+(\ell)})} p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{+(\ell)}) - p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)}), & \text{if } 0 \leq x_{k,n}^{(\ell)} \leq 2 \\ \infty, & \text{otherwise,} \end{cases} \tag{3.11}$$

where $\mathbf{x}^{+(\ell)}$ is the assignment vector after the modulation of user k on subchannel n is increased by one given the current $\mathbf{x}^{(\ell)}$. That is, $\mathbf{x}^{+(\ell)}$ is equal to $\mathbf{x}^{(\ell)}$ except that $x_{k,n}^{+(\ell)} =$

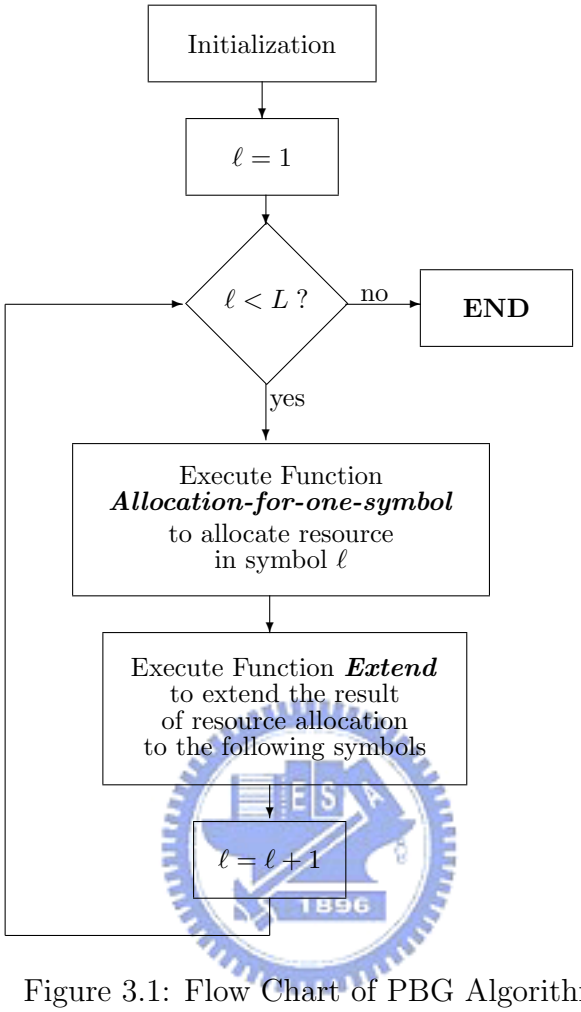


Figure 3.1: Flow Chart of PBG Algorithm

$x_{k,n}^{(\ell)} + 1$. Note that the increasing of 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 already in the subchannel is increased. The definition of the cost function in (3.11) 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. With the cost value defined above, the impact of inserting a user to a subchannel in SDMA scheme is identified by the additional power for maintaining the same modulation orders for the users that already in the subchannel.

Fig 3.1 shows the flow chart of the PBG algorithm. After The PBG algorithm accepts the information $\left\{\widehat{R}_k, 1 \le k \le K\right\}$ from the result of the dynamic priority adjustment scheme, the algorithm is performed as follows. The assignment variables, $x_{k,n}^{(\ell)}, \forall k, n, \ell$, are initially set to zero, which means that resource is not allocated to any user. Then, each user is given an instantaneous priority, and resource is allocated first to the users with high instantaneous priority. Denote β_k the instantaneous priority of user k, and it is initialized as the priority from dynamic priority adjustment scheme, i.e. $\beta_k = \widehat{R}_k$. 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 $\mathcal{N}_{free}^{(\ell)} = \{n | 1 \le n \le N\}$, for $1 \le \ell \le L$. The PBG algorithm then sequentially allocates resource for each OFDM symbol used for downlink transmission in a frame, from symbol 1 to symbol L. In each symbol, two functions are performed, function ***Allocation-for-one-symbol*** and function ***Extend***. The former iterative allocate resource to current OFDMA symbol, say symbol ℓ, while the latter try to extend the result of allocation to the following OFDMA symbols. 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* $\mathcal{N}_{free}^{(\ell)} = \{n | 1 \le n \le N\}$, *for* $1 \le \ell \le L$.

  for $\ell = 1 : L$ **do**

  Execute function ***Allocation-for-one-symbol***.

  Execute function ***Extend***.

  end for

Fig 3.2 shows the flow chart of the function ***Allocation-for-one-symbol***. In this function, 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^*),

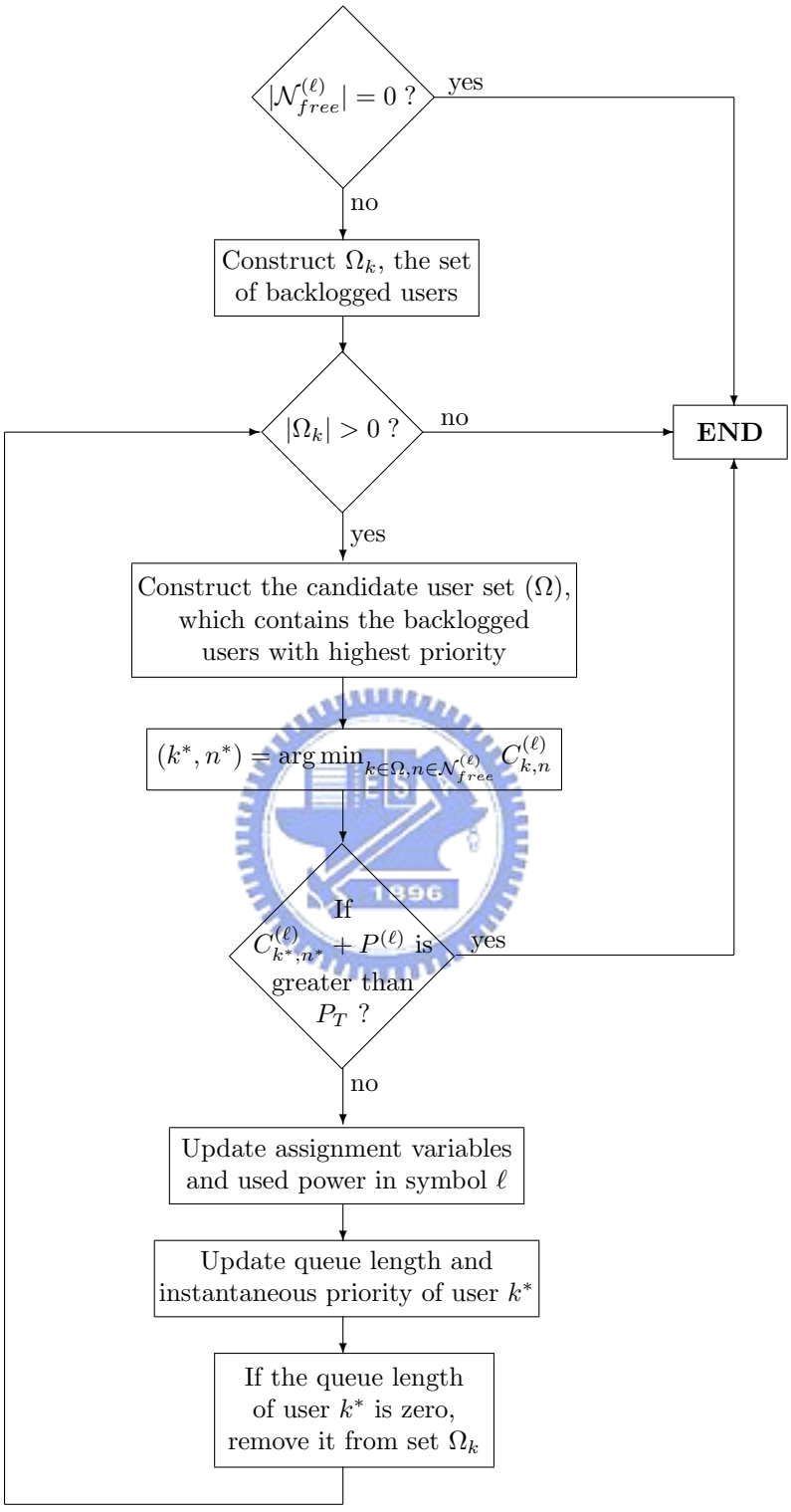
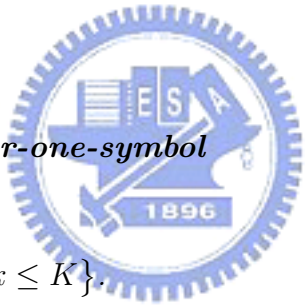


Figure 3.2: Flow Chart of Function ***Allocation-for-one-symbol***

is selected, in every iteration. The Ω contains the backlogged users with highest instantaneous priority. The (k^*, n^*) is selected from the users in the candidate user set and the free subchannels such that the cost value, $C^{(\ell)}_{k^*,n^*}$, 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. $x^{(\ell)}_{k^*,n^*} = x^{(\ell)}_{k^*,n^*} + 1$. Additional q bits are allocated to the selected user in each iteration, and thus the queue length of user k^*, R^B_k, 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 pseudocode of the function ***Allocation-for-one-symbol*** is given below.



- **Function:** ***Allocation-for-one-symbol***

  If $|\mathcal{N}^{(\ell)}_{free}| = 0$, *then* **end**.

  Set $\Omega_K = \{k | R^B_k > 0, 1 \le k \le K\}$.

  while $|\Omega_K| > 0$

    Set $\Omega = \{k | \beta_k = \beta_{max}, k \in \Omega_K\}$, *where* $\beta_{max} = \max_{k \in \Omega_K} \beta_k$.

    Find $(k^*, n^*) = \arg\min_{k \in \Omega, n \in \mathcal{N}^{(\ell)}_{free}} C^{(\ell)}_{k,n}$.

    If $C^{(\ell)}_{k^*,n^*} + P^{(\ell)} > P_T$, *then* **end**.

    Set $x^{(\ell)}_{k^*,n^*} = x^{(\ell)}_{k^*,n^*} + 1$ *and* $P^{(\ell)} = P^{(\ell)} + C^{(\ell)}_{k^*,n^*}$.

    Set $R^B_{k^*} = \max(R^B_{k^*} - q, 0)$ *and* $\beta_{k^*} = \max(\beta_{k^*} - q, 0)$.

    If $R^B_{k^*} = 0$, *then* $\Omega_K = \Omega_K - \{k^*\}$.

  end while

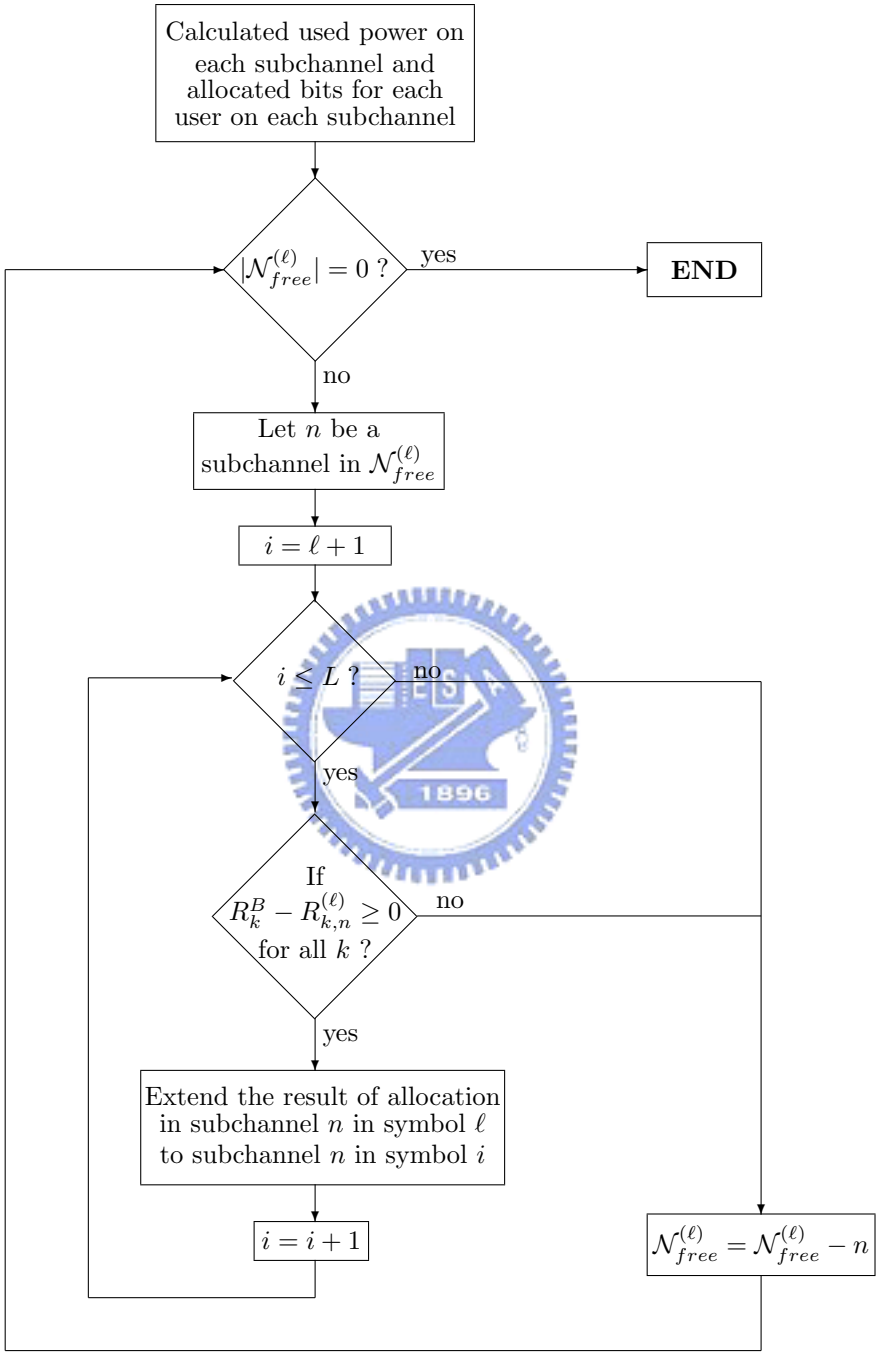


Figure 3.3: Flow Chart of Function ***Extend***

Fig 3.3 shows the flow chart of the function ***Extend***. In this function, 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. For each free subchannel in the ℓth OFDMA symbol, the same assignment variables would be given to the following OFDMA symbol, i.e. $x_{k,n}^{(i)} = x_{k,n}^{(\ell)}, \forall k$ for some i greater than ℓ, if the additional transmission bits to user does not exceed its queue length. The assignment variables, instantaneous priority, queue length, used power, and the set of free subchannels are updated if the extension is performed. The pseudocode of the function ***Extend*** is given below.

- **Function:** ***Extend***

  Calculate $P_n^{(\ell)} = \sum_k p_{k,n}^{(\ell)}(BER_k^*, \mathbf{x}^{(\ell)})$, $\forall n$.

  Calculate $R_{k,n}^{(\ell)} = q \times x_{k,n}^{(\ell)}$, $\forall k, n$.

  for all n *in* $\mathcal{N}_{free}^{(\ell)}$ **do**

  $i = \ell + 1$.

  while $R_k^B - R_{k,n}^{(\ell)} \geq 0, \forall k$ *and* $i \leq L$ **do**

  Set $x_{k,n}^{(i)} = x_{k,n}^{(\ell)}$ $(\forall k)$ *and* $P^{(i)} = P^{(i)} + P_n^{(\ell)}$.

  Set $R_k^B = R_k^B - R_{k,n}^{(\ell)}$ *and* $\beta_k = \max(\beta_k - R_{k,n}^{(\ell)}, 0)$, *for all* k.

  let $\mathcal{N}_{free}^{(i)} = \mathcal{N}_{free}^{(i)} - \{n\}$ *and* $i = i + 1$.

  end while

  end for

The proposed ARRA algorithm is summarized as follows. At the beginning of each frame, the base station begin to allocate resource to all users. The dynamic priority adjustment scheme is performed and the result, $\widehat{R}_k$, is transferred to the PBG algorithm. The base station then operates the PBG algorithm to find the solution of ARRA algorithm, $\mathbf{x}^{(\ell)}, 1 \leq \ell \leq L$.

After that, the subchannel allocation, modulation order assignment, power allocation, and transmit beamforming can be obtained from the solution. The packet of each user is then transmitted by the base station. Finally, the queue state of each user is updated according to the result of resource allocation and the arrival of new packets.

3.2 Conventional Algorithms

For comparison, the following three conventional RRA algorithms are described here.

3.2.1 Linkgain-based (LB) resource allocation

The resource is allocated to users according to the users' CSI in LB algorithm. The LB algorithm is an simple algorithm to select the user with better channel for throughput maximization. For each subchannel, the first Q users with best channel quality are selected and the highest possible modulation order without violating the total power constraint is set for the users. The algorithm sequentially allocates subchannels to users until the the remaining power cannot support the lowest possible modulation order for the selected users. The same procedure is repeated for each symbol in a frame.

The algorithm is modified from [1], which was based on OFDM system without multiple antenna. In [1], the algorithm selected the best one user is selected for each subchannel, but to fit the architecture of multiple antenna, the best Q users are selected in LB scheme. Note that in above statements, only the users with nonempty buffer are considered. The algorithm would not select a user that does not has data in its buffer even if the channel quality of the user is good.

3.2.2 Multi-antenna Multi-user Maximum Sum Rate (MMSR)

In MMSR algorithm [10], the resource allocation is taken on two step approach, *user clustering* and *space-frequency bit allocation*. The algorithm execute this two steps for each

OFDMA symbol in the current frame. In first step, the user with best channel quality is first selected for each subchannel. After that, users are added to a subchannel based on an scalar product which measuring the spatial correlation between the users already in the subchannel and the other users. The scalar product of user k and k' in subchannel n for the ℓth OFDMA symbol is given by,

$$c_{k,k'} = \sum_{i \in \Psi_n} \frac{|\mathbf{h}_{k,i} \mathbf{h}_{k',i}^{\dagger}|}{\|\mathbf{h}_{k,i}\| \, \|\mathbf{h}_{k',i}\|}, \tag{3.12}$$

where $\mathbf{h}_{k,i}^{\dagger}$ denote the complex conjugate transpose of $\mathbf{h}_{k,i}$ Q users are grouped in each subchannel and only the users with nonempty buffer may be selected. The procedure of the step *user clustering* while selecting the users for OFDMA symbol ℓ is summarized as,

- ***User Clustering in MMSR***

  Set $n = 0$.

  Let $\mathcal{K}_n^{(\ell)}$ *has the best user for each subchannel* n.

  for $n = 1 : N$ **do**

    for $q = 2 : Q$ **do**

      $k^* = \arg\min_k \max_{k' \in \mathcal{K}_n^{(\ell)}} c_{k,k'}$.

      Add user k^* *to* $\mathcal{K}_n^{(\ell)}$.

    end for

  end for

The modulation order is then determined by the step *space-frequency bit allocation*. Briefly, the step first try to server all the users selected in first step with highest modulation order. If the power constraint is not fulfilled, the algorithm decides which user among all subchannels should reduce its modulation order. The algorithm select the user that save

the most power if its modulation order is reduced. The BS reduces the number of bits of the selected user the transmission power is recalculated. The algorithm finishes when the power constraint is fulfilled. The detailed procedure is given in [10].

3.2.3 Truncated Generalized Processor Sharing (TGPS)

In [12], the GPS scheduling is used in an OFDMA system. The algorithm is modified here to fit the architecture of multiple antenna. In TGPS algorithm, resource is allocated to users based on a predefined weights. Also, the adaptive modulation is not used in TGPS algorithm and hence all users transmit data with a predefined modulation order. The algorithm first determined the number of required transmission bits for each user in current OFDMA symbol based on predefined weights for all users. If the system can transmit $R_{\text{effective}}$ bits in current OFDMA symbol, the required transmission bits for user k, denoted by $R_{\text{TGPS},k}$, is given by

$$R_{\text{TGPS},k} = \left\lfloor \frac{\eta_k}{\sum_{u\in\Omega_k}\eta_u} R_{\text{effective}} \right\rfloor, \tag{3.13}$$

where η_k is the weight of user k and Ω_k is the set of backlogged users. And the quantization error of user k, denoted by $\Delta R_{\text{TGPS},k}$, is given by

$$\Delta R_{\text{TGPS},k} = \frac{\eta_k}{\sum_{u\in\Omega_k}\eta_u} R_{\text{effective}} - R_{\text{TGPS},k}. \tag{3.14}$$

The left bits due to the quantization in (3.13), is allocated by a compensation algorithm, where the users with large cumulative quantization error get the left bits.

After that, a simplified power and subchannel allocation algorithm is used, where it try to satisfy the required transmission bits for each user. Resource is allocated to the users with high SNR required first in the simplified power and subchannel allocation algorithm.

The BS then calculated the total transmission power and compares it with the total system power, P_T. If the total transmission power is less than the total system power, the resource allocation is complete. Otherwise the $R_{\text{effective}}$ is decreased by one unit, which is the number of transmission bits of a user over one subchannel when the predefined modulation scheme is used, and the above steps is repeated. The algorithm finishes when each user can transmit at least its required number of transmission bits and the power constraint is fulfilled. The detailed procedure is given in [12].

Chapter 4

Simulation Results and Discussion

In the simulation, the system-level OFDMA/SDMA downlink environments are set to be compatible with IEEE 802.16 standard [13]. These system-level parameters are listed in Table 4.1, and scalable parameters in physical layer are configured according to the suggested values in [19]. The OFDMA/SDMA system is based on 5 MHz bandwidth and the frame duration is 2 ms. The number of subcarriers is equal to the FFT size, 512, but only 384 subcarriers are used for data transmission, while the others are used for pilot channel or guard channel.

The wireless fading channel consists of large-scale fading and small-scale fading. The large-scale fading comes from free space degrading and shadowing effect, while the small-scale fading is due to multipath reflection [20]. 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 [15]. Also, the log-normal shadowing with zero mean and standard deviation of 8 dB is assumed. The multipath channel for each antenna has six taps of Rayleigh-faded paths and the power delay profile follows exponential decay rule. Different users have independent channels but with the same statistics. The channel is assumed to be fixed within a frame and varies independently from frame to frame.

Parameters	**Values**
Cell size	1600m
Number of antenna at BS (Q)	3
Frame duration	2ms
System bandwidth	5 MHz
Sampling frequency	5.714 MHz
FFT size	512
Subcarrier frequency spacing	11.16 kHz
OFDMA symbol duration	100.8 μs
Useful symbol time	89.6 μs
Guard time	11.2 μs
Number of data subcarriers	384
Number of subchannels (N)	8
Number of data subcarriers per subchannel (b)	48
Basic allocation unit (q)	96 bits
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
Propagation model	$128.1 + 37.6 \log R$ dB
Fast fading model	Rayleigh distribution
Standard deviation of slow fading	8 dB

Table 4.1: System-Level Configuration

Component	Distribution	Parameters
Inter-arrival time between each video frame	Deterministic	100ms
Number of packets (slices) in a video frame	Deterministic	8
Packet size	Truncated Pareto (Mean = 100 bytes, Max = 250 bytes)	K = 40 bytes, $\alpha = 1.2$
Inter arrival time between packets (slices) in a frame	Truncated Pareto (Mean = 6ms, Max = 12.5ms)	K = 2.5ms, $\alpha = 1.2$

Table 4.2: Video Streaming Traffic Model Parameters

4.1 Source Traffic Model and QoS requirements

The QoS requirements and detailed parameters for the four traffic models are given in this section. Each voice user traffic is modeled as an ON-OFF model, in which the length of ON period and OFF period follows the exponential distribution. The mean of ON period for voice traffic is 1 second while the mean of OFF period is 1.35 seconds. During ON period, the mean data rate is 12.2 kbps. The video Streaming consists of a sequence of video frames which are emitted regularly with interval 100ms. Every video frame in the streaming is composed of eight slices, and each slice is corresponding to a single packet. The video traffic model parameters are defined in Table 4.2 [15], where the source data rate is 64 kbps.

Parameters for HTTP traffic model are defined in Table 4.3 [15]. Note that the packets of HTTP traffic used a maximum transmission unit of 1500 bytes. As to FTP traffic model, the size of each file is in truncated lognormal distribution with mean 2Mbytes, standard deviation 0.722 Mbytes, and a maximum value 5 Mbytes. The interval between files is exponentially distributed with mean 180 seconds. Table 4.4 lists the QoS requirements of

Component	Distribution	Parameters
Main object size	Truncated Lognormal	Mean = 10710 bytes, Std. dev. = 25032 bytes, Max = 2 Mbytes, Min = 100 bytes
Embedded object size	Truncated Lognormal	Mean = 7758 bytes, Std. dev. = 126168 bytes Max = 2 Mbytes, Min = 50 bytes
Number of embedded objects per page	Truncated Pareto	Mean = 5.64, Max = 53
Inter arrival time between each page	Exponential	Mean = 30 second
Packet size	Deterministic	Chop from objects with size 1500 bytes
Packet inter arrival time	Exponential	Mean = 0.13 second

Table 4.3: HTTP Traffic Model Parameters

each traffic type. The required BER of NRT services or BE service is larger than that of RT services for accurate data transmission. In addition, the minimum required transmission rate of HTTP users (in NRT services) is slightly larger the arrival rate of HTTP traffic in page download. This means transmission rate for download of an web page is guaranteed and hence the response time of a web browser is small.

4.2 Performance Evaluation

In this 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 as the ratio of the total average arrival rate of all users over the system maximum transmission rate. The maximum transmission rate is achieved when Q users are multiplexed for each subchannel

Traffic Type	Requirement	Value
voice (RT)	required BER	10^{-3}
	maximum delay tolerance	40ms (20 frame)
	maximum allowable dropping ratio	1%
video (RT)	required BER	10^{-4}
	maximum delay tolerance	10ms (5 frame)
	maximum allowable dropping ratio	1%
HTTP (NRT)	required BER	10^{-6}
	minimum required transmission rate	100 kbps (200 bits per frame)
FTP (BE)	required BER	10^{-6}

Table 4.4: The QoS Requirement of each traffic type

and the highest modulation order is used for all users. It is equal to 27.648 Mbps in the simulation environment of this paper. Note that the average arrival rate of voice, video, HTTP, and FTP user is equal to 5.2 kbps, 64 kbps, 14.5 kbps, and 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 proposed ARRA algorithm, LB algorithm, MMSR algorithm, and TGPS algorithm are simulated in the simulation. The modulation scheme for TGPS is fixed at 16-QAM since the performance of TGPS is best while using this modulation level. The predefined weight for TGPS is set to 10, 5, and 1 for RT, NRT, and BE service, respectively. The following performance measures are considered: (i) system throughput, (ii) packet dropping rate of RT users, (iii) mean packet delay of RT users, (iv) average transmission rate of NRT users, (v) Guaranteed ratio of NRT users, which is defined as the ratio of the number of the NRT users whose average transmission rate are greater than the minimum required transmission rate over total NRT users, and (vi) average transmission rate of BE users.

Fig. 4.1 shows the system throughput versus the traffic load. When the traffic load is smaller than 0.2, all the four algorithms have the same throughput. However, the system

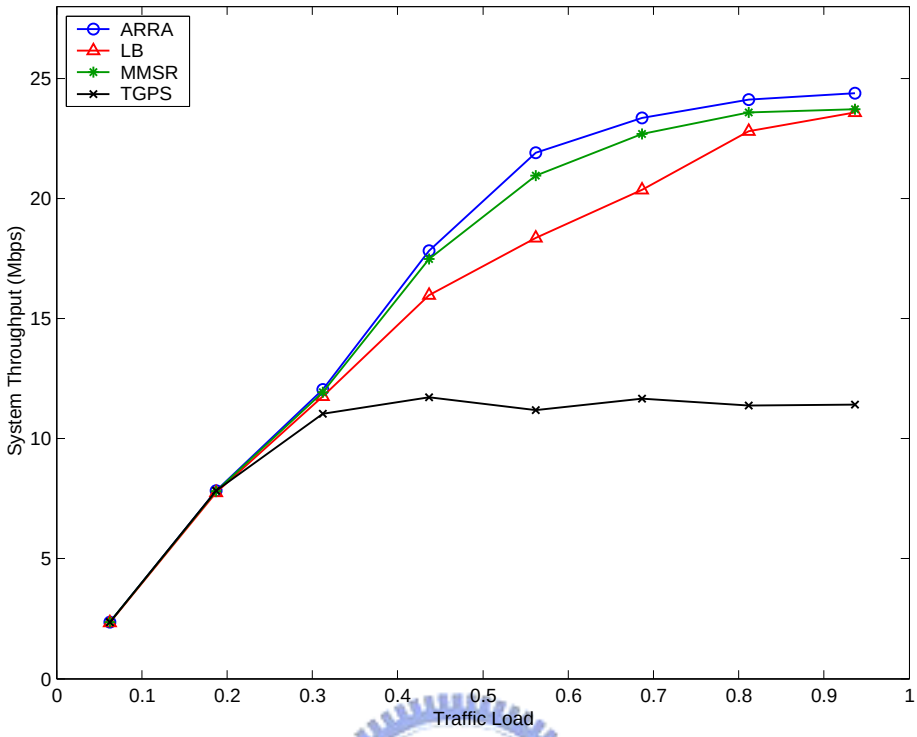
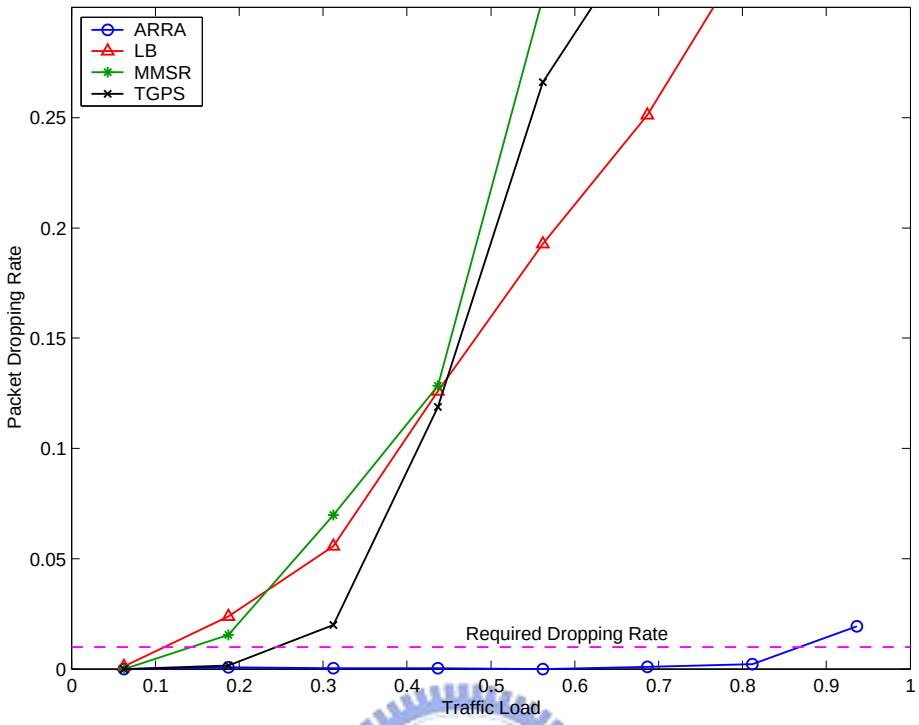


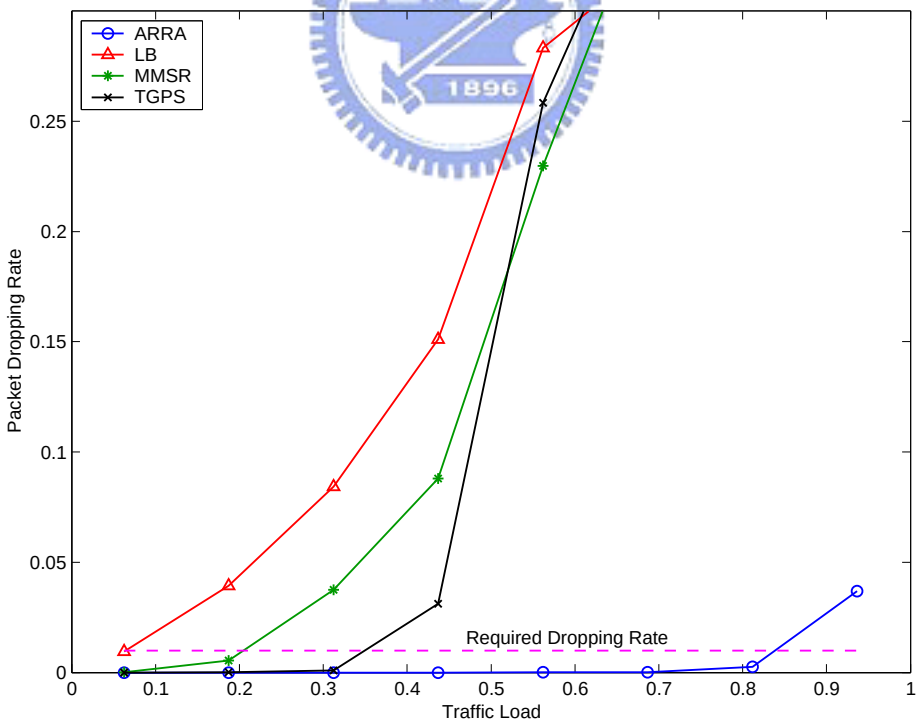
Figure 4.1: System Throughput

throughput of ARRA, LB, or MMSR is much higher than that of TGPS when the traffic load is further increased. This is because that the TGPS uses simplified algorithm for subchannel allocation and multiuser diversity is not well exploited. Also, adaptive modulation is not enable in TGPS and make less efficient usage of power. The ARRA algorithm is a little bit better than LB algorithm when the traffic load is greater than 0.4 for the reason that the optimal user grouping in space domain is not considered in LB algorithm. The system throughput of proposed ARRA algorithm is closed to that of MMSR since both of this two schemes takes throughput maximization as one of design objective. Note that the throughput ARRA algorithm can reach 24 Mbps, which is very closed to the system's maximum throughput, 27.648 Mbps. Hence, the ARRA algorithm improve the system throughput by taking multiuser diversity and space domain correlation between users in its design.

Fig. 4.2(a) and 4.2(b) depict the packet dropping rate of voice users and the packet



(a) Packet Dropping Rate of Voice Users

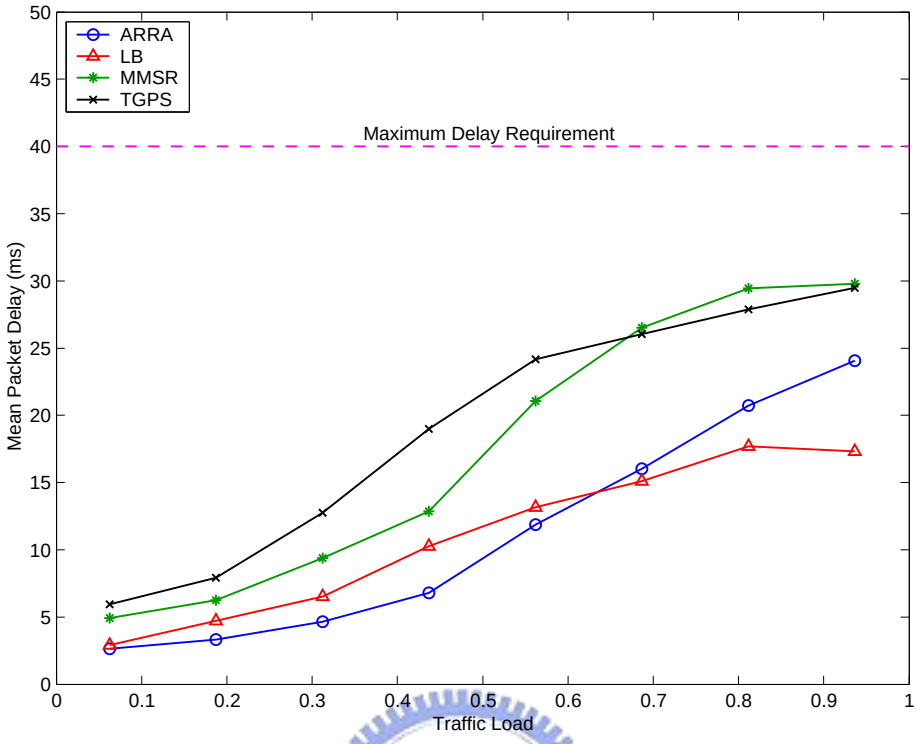


(b) Packet Dropping Rate of Video Users

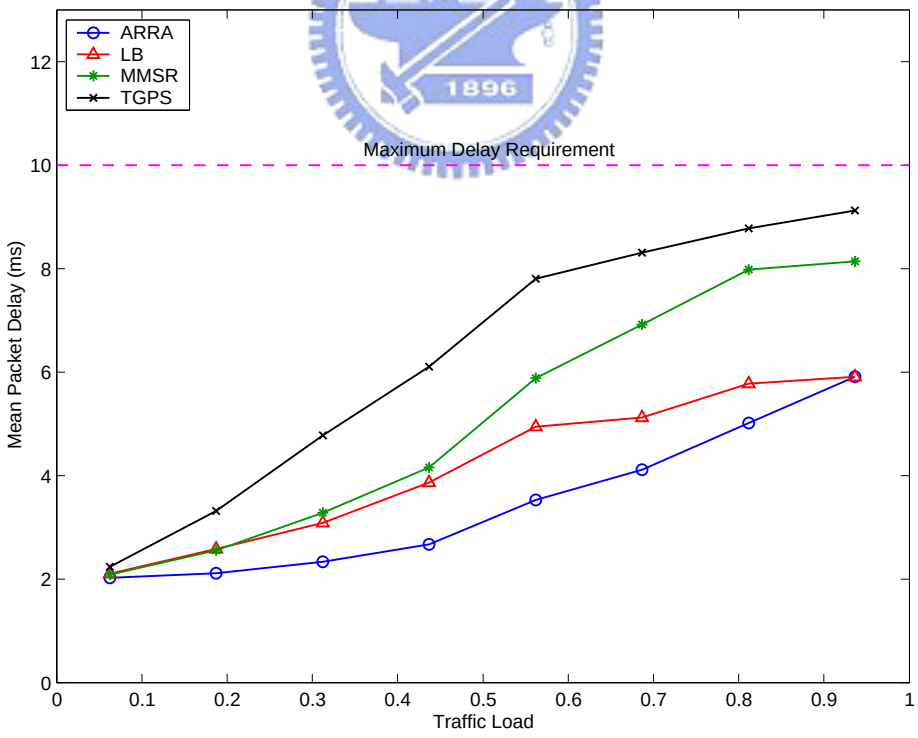
Figure 4.2: Packet Dropping Rate of RT Users

dropping rate of video users, respectively. These figures show that the packet dropping ratio of the ARRA algorithm is almost zero until the traffic load is greater than 0.8, while that of LB or MMSR increases rapidly with the traffic load and violates the maximum allowable dropping rate at very light traffic load. The reason is that the LB or MMSR algorithm does not consider the QoS requirement of maximum delay tolerance for the RT users. The dropping rate of TGPS algorithm is quite small when the traffic load is smaller than 0.3 since the TGPS algorithm bias the RT users by giving them large weight. However, the dropping rate is too large when the TGPS algorithm reaches its maximum capacity, which is about 11 Mbps as shows in Fig 4.1. In this case, the arrival rate is lager than the capacity that the TGPS algorithm can support and hence the dropping rate is larger than the maximum allowable dropping rate even through the TGPS algorithm bias the RT users. On 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 rate is very small.

Fig. 4.3(a) and 4.3(b) show the mean delay of voice users and the mean delay of video users, respectively. For all the three algorithms, the mean packet delay increases with the traffic load. This phenomenon is expected since less resource is provided to individual RT user for a large number of users. The TGPS algorithm usually has largest packet delay because that most packets require more than one frame for completing its transmission. Although the TGPS allocates more bits to the RT users in each frame than that to NRT users or BE users, these bits is evenly distributed to all RT users. The number of bits allocate to a RT user is usually smaller than the packet length of its HOL packet and requires more frames for completing the transmission of the HOL packet. On the other hand, the extend function in ARRA algorithm makes the selected RT users to transmit over several OFDMA symbol. Therefore, once a RT user is selected by the algorithm, it can complete the transmission of



(a) Mean Packet Delay of Voice Users



(b) Mean Packet Delay of Video Users

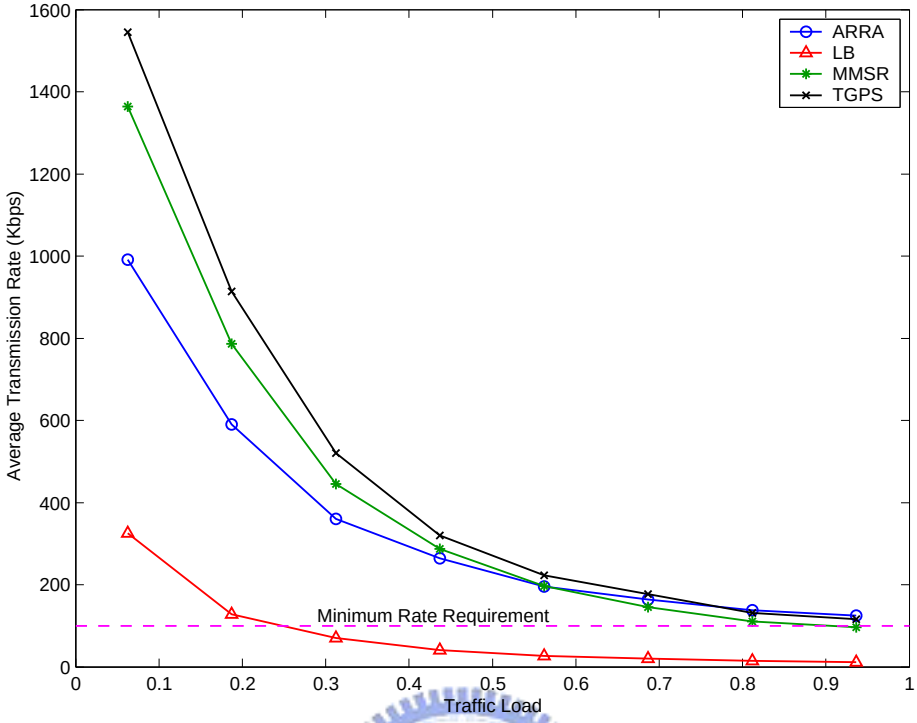
Figure 4.3: Mean Packet Delay of RT Users

its HOL packet with large probability, which results in low packet delay.

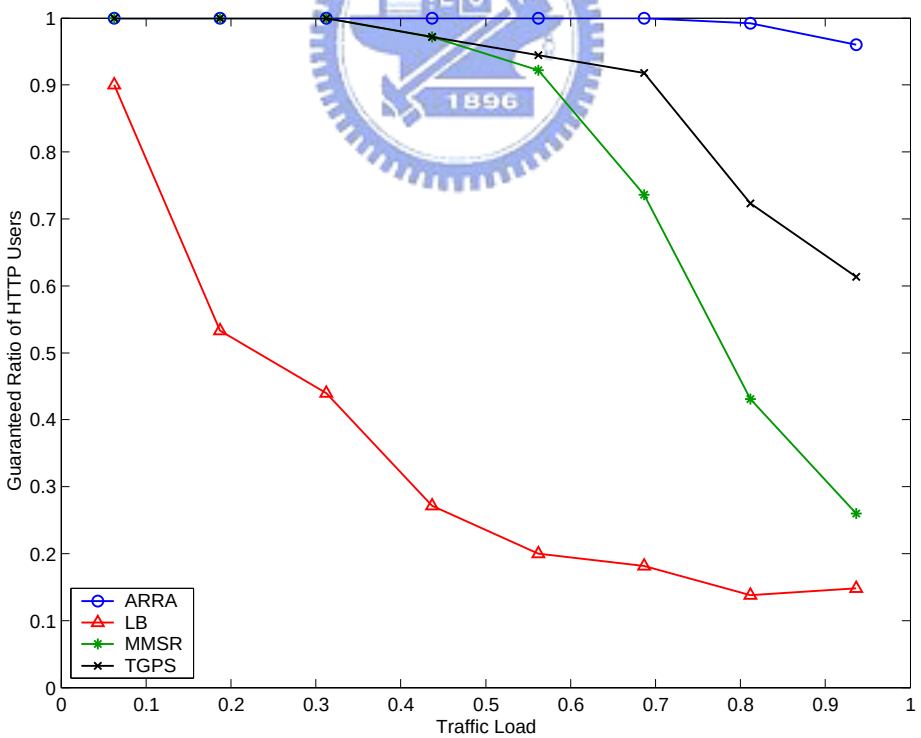
In most case, the delay of the ARRA algorithm is lower than that of TGPS, LB, or MMSR algorithm for the reason that the priority of RT users is usually higher than other users in ARRA algorithm. The mean packet delay of voice users of LB algorithm is less than that of ARRA algorithm when the traffic load is higher than 0.7. However, at that kind of traffic load, the packet dropping ration of voice users of LB algorithm is much higher than that of the ARRA algorithm. This means that the LB algorithm can only provide very low delay to a certain of voice users and results in a lower mean packet delay than the ARRA algorithm, but the dropping rate of the other voice users is too high. On the contrary, the ARRA can give almost all voice users low dropping rate with the satisfaction of the increasing of mean delay. From Fig. 4.2 and Fig. 4.3, we can conclude that the ARRA algorithm outperforms the conventional methods in the satisfaction of the QoS requirements of RT users.

Fig. 4.4(a) and 4.4(b) illustrate the average transmission rate of HTTP users and the guaranteed ratio of HTTP users, respectively. For the ARRA algorithms, the average transmission rate decreases as the traffic load increases, but the minimum required transmission rate for NRT users is guaranteed. The MMSR and TGPS algorithms show the same phenomenon and the average transmission rate is higher than the transmission rate in the ARRA algorithm at low traffic load. However, for the LB algorithm the average transmission rate is smaller than the minimum required transmission rate when the traffic load is larger than 0.2. This is because that the only guarantee the NRT users with good channel quality and thus the transmission rate of the other NRT users is very small, which resulted in a low average transmission rate.

The ARRA algorithm guarantees the minimum transmission rate of each NRT users by giving high priority to the NRT users with transmission rate lower than minimum required transmission rate and thus the average transmission rate of all NRT users is guaranteed.



(a) Average Transmission Rate of HTTP Users



(b) Guaranteed Ratio of HTTP Users

Figure 4.4: Performance of NRT Users

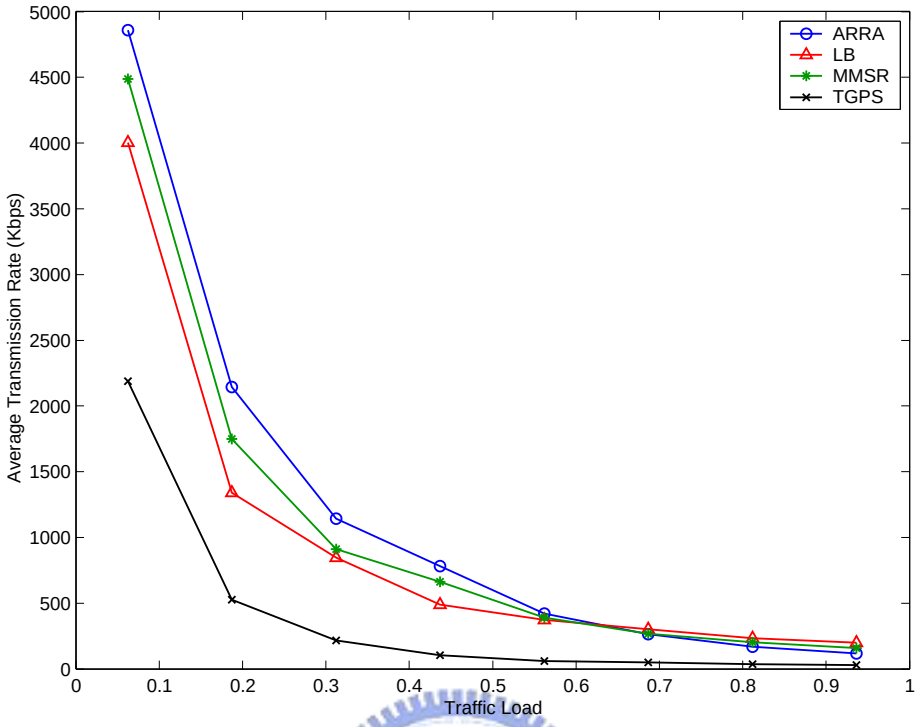


Figure 4.5: Average Transmission Rate of FTP Users

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 of HTTP users 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. From Fig. 4.4(a) and 4.4(b), we can see that, in the ARRA algorithm, the average transmission rate of all NRT users is acceptable and each NRT user is guaranteed with a minimum transmission rate.

Fig. 4.5 shows the average transmission rate of FTP users. Although the ARRA algorithm does not guarantee the transmission rate of the users in BE service class. The transmission rate of the ARRA algorithm is still higher than other algorithms when the traffic load is

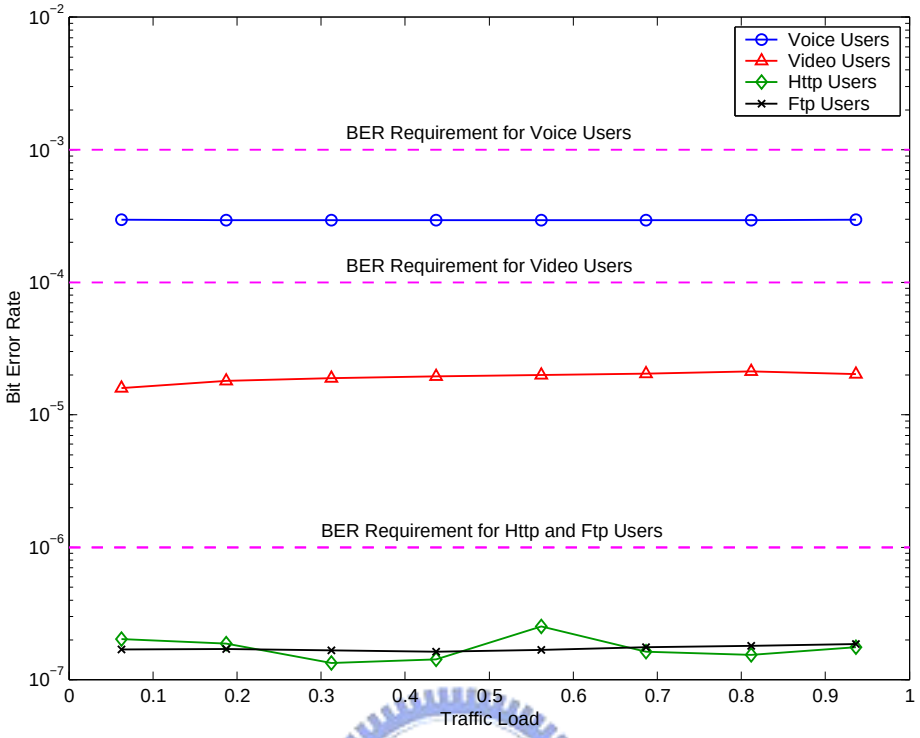


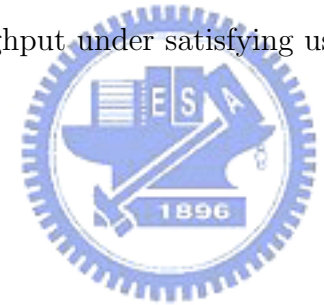
Figure 4.6: BER of the ARRA algorithm

less than 0.7. This is because that the goal of throughput maximization in ARRA algorithm makes the transmission rate of each user as high as possible. The transmission rate of FTP users in TGPS algorithm is very small since the TGPS algorithm gives less resource to FTP users by setting the weight of FTP user smaller than that of other users. The transmission rate of ARRA algorithm is lower than LB or MMSR algorithm in high traffic load for the reason that the ARRA algorithm has to guarantee the QoS requirements in other service classes. We believe that this is an worthwhile tradeoff since the ARRA algorithm is much better than the other algorithms in the satisfaction of QoS requirements, as show in above figures.

Fig. 4.6 shows the BER of the users in each traffic type for the ARRA algorithm. It can be seen that the BER is below the the required BER for all the four traffic type. This is because that once a user is transmitting, the allocated power to the user is always make the

receive SINR above a threshold, which can be obtained from the user's modulation order and required BER. The detail setting of power allocation can be referred in section 2.5. Therefore, the QoS requirement of BER is guaranteed in the ARRA algorithm.

From the above simulation results, the ARRA algorithm obviously outperforms the conventional algorithms in terms of system throughput and the satisfaction of the QoS requirements for both RT and NRT services. The performance of the ARRA algorithm in BE service is also acceptable. It is because the ARRA algorithm can give high priority to the user that really need more resource and the PBG algorithm can efficiently use power to maximum system throughput. NRT users with low average transmission rate and RT users with large packet delay can obtain the resource earlier. When no urgent users, the ARRA algorithm take throughput maximization as its objective. As a result, the ARRA algorithm can achieve better system throughput under satisfying users' QoS requirements.



Chapter 5

Conclusions

In this thesis, an adaptive radio resource allocation (ARRA) algorithm is proposed for downlink OFDMA/SDMA systems. The subchannel allocation, modulation order assignment, power allocation, and beamforming are controlled by the ARRA algorithm. The goals of ARRA algorithm are QoS satisfaction and throughput maximization. In addition, multiple service classes, which include RT, NRT and BE services, are considered. The proposed ARRA algorithm gives high priority to the urgent users at the current frame and dynamically adjust the priority of users frame by frame. An iterative algorithm, named PBG algorithm, is also presented to efficiently allocate the resource based on a cost value.

In the simulation results and discussion, the ARRA algorithm is compared with conventional algorithms, LB, MMSR, and TGPS algorithm. From the results, we can conclude that the ARRA algorithm outperforms the conventional algorithms in terms of system throughput and the satisfaction extent of QoS requirements. In the ARRA algorithm, NRT users with low average transmission rate and RT users with large packet delay can obtain the resource earlier. In addition, the ARRA algorithm take throughput maximization as its objective when there are no urgent users. As a result, the ARRA algorithm can achieve better system throughput under satisfying users' QoS requirements.

The ARRA algorithm can cooperate with other radio resource management techniques to be implemented on real system. Accompanying with a well-designed call admission control

(CAC) scheme, the ARRA algorithm would be operated on a suitable traffic load such that the users' QoS requirements can be easily fulfilled. In the multi-cell environment, the ARRA algorithm can be accompanied with a dynamic channel allocation (DCA) to mitigate the inter-cell interference.

Bibliography

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 3GPP TR 25.892, "Feasibility study for OFDM for UTRAN enhancement," 3rd Generation Partnership Project, Tech. Rep., 2004-06.

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

[19] H. Yaghoobi, "Scalable OFDMA physical layer in IEEE 802.16 WirelessMAN," *Intel Techonology Journal*, Volume 8, Issue 3, 2004.

[20] T. S. Rappaport, *Wireless communications: principles and practice*. Upper Saddle River, NJ: Prentice Hall, 1996.

Vita

Chun-Fan, Tsai was born in Miaoli, Taiwan. He received B.E. and M.E. degree in department of communication engineering from Nation Chiao-Tung University, Hsinchu, Taiwan, in 2004 and 2006, respectively. His research interests include radio resource managment and wireless communication.