Proportional Fair Scheduling Algorithm in OFDMA-Based Wireless Systems with QoS Constraints

Tolga Girici, Chenxi Zhu, Jonathan R. Agre, and Anthony Ephremides

Abstract: **In this work we consider the problem of downlink resource allocation for proportional fairness of long term received rates of data users and quality of service for real time sessions in an OFDMA-based wireless system. The base station allocates available power and subchannels to individual users based on long term average received rates, quality of service (QoS) based rate constraints and channel conditions. We formulate and solve a joint bandwidth and power optimization problem, solving which provides a performance improvement with respect to existing resource allocation algorithms. We propose schemes for flat as well as frequency selective fading cases. Numerical evaluation results show that the proposed method provides better QoS to voice and video sessions while providing more and fair rates to data users in comparison with existing schemes.**

Index Terms: **Heterogeneous traffic, IEEE 802.16, orthogonal frequency division multiple access (OFDMA), quality of service (QoS), proportional fairness, resource allocation, WiMax.**

I. INTRODUCTION

Increasing number of users demanding wireless Internet access and a growing number of wireless applications require high speed transmission and efficient utilization of system resources such as power and bandwidth. Recently, technologies like WiMax (based on IEEE 802.16 standard) [1] and Long Term Evolution of 3GPP (LTE) [2] are developed to address these challenges. Orthogonal frequency division multiplexing (OFDM), a multicarrier transmission technique, is the preferred transmission technology in next generation broadband wireless access networks. It is based on a large number of orthogonal subcarriers, each working at a different frequency. OFDM is originally proposed to combat intersymbol interference and frequency selective fading. However, it also has a potential for a multiple access scheme, where the subcarriers are shared among the competing users. Within orthogonal frequency division multiplexing access (OFDMA) framework, the resource allocated to the users comes in three dimensions: Time slots, frequency, and power. This requires the scheduler to operate with higher degree of freedom and more flexibility, and potentially higher multiplexing capacity. This also increases the dimensionality of the resource allocation problem and makes the problem more involved. We plan to develop scheduling algorithms fully taking advantage of the degree of freedom inherent to OFDMA system. Our goal is to find multicarrier proportional fair schemes that also satisfy heterogeneous stability and delay requirements.

There are three main issues that need to be considered in multiple access resource allocation. The first one is *spectral efficiency*, which means achieving maximum total throughput with available bandwidth and power. In time division multiple access (TDMA) transmission, it is achieved by always allowing the user with the best channel to transmit [3]. In OFDM multiple subcarriers possibly experience different fading levels, which makes spectral efficiency a more complex problem. The second issue is fairness. In [3], it is studied that if the channel conditions are independent and identically distributed (i.i.d.), all users eventually will get the same service, hence fairness is maintained [4]. On the other hand if the distance attenuations of users are different then some users will definitely get more service, and some others won't get any service. Therefore scheduling algorithms have to be proposed to provide fairness among nodes. The third important issue is satisfying quality of service (QoS) requirements. An example for QoS requirements can be bounds on delay or short term received rate for real time applications.

In broadband wireless access systems proportional fairness maintains a good tradeoff between spectral efficiency and fairness. The proportional fair scheduler is proposed by Jalali *et al.* in [1] in the context of high data rate (HDR) system. This system is originally designed for data applications(e.g., FTP and Internet). Basically, the system is a TDMA system where a user is scheduled to transmit at each time slot. A proportional fair schedule is such that any positive change of a user in rate allocation must result in a negative overall change in the system. Proportional fair resource allocation also corresponds to an allocation that maximizes the sum of logarithms of received rates [5].

Proportional fair scheduling was recently studied for multicarrier systems in [5], [6], [7], and [8]. In [5], proportional fairness formulation is extended from single to multiple channel systems. However it did not include power control and no algorithm was proposed to find the optimum bandwidth allocation. In [6], [7], and [9], proportional fair scheduling is addressed for a single time instant, that is, their objective was to maximize the sum of logarithms of instantaneous rates, rather than the long term received rates. The recent paper [8] proposes an algorithm for proportional fairness of long term rates in OFDM, however it doesn't consider power control. Besides in all of these works supporting real time traffic was not addressed. The scheduling rules do not apply sufficiently to different QoS requirements and heterogeneous traffic.

Most of the previous work on resource allocation for heterogeneous traffic requirements followed and extended the ap-

Manuscript received April 04, 2008; approved for publication by Hlaing Minn, Division II Editor, September 21, 2009.

T. Girici is with the Dept. of Electrical and Electronics Eng. of TOBB University of Economics and Technology, Ankara, Turkey, email: tgirici@etu.edu.tr.

C. Zhu and J. R. Agre are with the Fujitsu Labs of America at College Park, 8400 Baltimore Ave., Suite 302, College Park, MD, 20740, USA, email: {chenxi.zhu, jonathan.agre}@us.fujitsu.com.

A. Ephremides is with the Department of Electrical Engineering and Institute of Systems Research at University of Maryland, College Park, MD, 20740, USA, email: etony@umd.edu.

proach in [10]. In single channel systems, largest weighted delay first (LWDF) proposed in [10] is shown to be throughput optimal. According to this scheme at each time instant , the user that maximizes a combination of head-of-line packet delay, averaged received service and current achievable rate is served. This scheme is extended for OFDMA-based multichannel systems in [11], [12], and [13]. In [11] average delay (that can be estimated by using average queue size and arrival rate) is used instead of head-of-line delay. Using head-of-line delay is shown in [12] to perform better than using average delay. All these proposed schemes assume uniform power. In fact, power control can be useful in improving the performance of these schemes. Finally, in [13] power control is considered, however, power allocated to a user is proportional to its number of subchannels, therefore power control is still not fully exploited. These schemes are explained in more detail in Section III. Our main contribution in this work is formulating a joint bandwidth/power optimization framework that takes more advantage of power control.

OFDMA based resource allocation has been studied also without the proportional fairness objective in [14]–[19]. The works [14] and [20] propose subcarrier and bit allocation algorithms that satisfy rate requirements of users with minimum total power. The authors in [15], [17], [18], and [19] address maximizing total and weighted throughput subject to power and subcarrier constraints and do not address real time traffic. The work in [16] introduces a proportional rate constraint, where the rates of individual users have to be in certain constant proportions in order to maintain fairness. However, this also doesn't guarantee any short or long term transmission rates.

The system considered in this work is motivated by the recent IEEE 802.16 standard that defines the air interface and medium access control (MAC) specifications for wireless metropolitan area networks. Such networks intend to provide high speed voice, data and on demand video streaming services for end users. IEEE 802.16 standard is often referred to as WiMax and it provides substantially higher rates than cellular networks. Besides it eliminates the costly infrastructure to deploy cables, therefore it is becoming an alternative to cabled networks, such as fiber optic and DSL systems [1].

In IEEE 802.16e, in order to ease the resource allocation process the subcarriers are grouped into subchannels. There are two main classes of subchannelization methods. The first class is adaptive modulation and coding (AMC). In this method a number of carriers adjacent on the frequency spectrum are grouped into a *band AMC subchannel*. In a multipath fading channel different subchannels experience different levels of fading. Achievable rates can be maximized by adjusting the modulation and coding rate according to the fading level for each subchannel. The second class includes partial use of subchannels (PUSC) and full use of subchannels (FUSC). They are diversity permutation schemes that distribute the sub-carriers of a subchannel pseudo-randomly in a wide frequency band. They provide frequency diversity and inter-cell interference averaging. This relieves the performance degradation due to fast fading mobile environments. PUSC is the default mode of subchannelization and is more suitable for mobile users than AMC. Here, because of the averaging effect, we can also assume that each subchannel experiences the same fading with respect to a user. This decreases the complexity of resource allocation algorithms because the problem becomes *how many subchannels* instead of *which ones*. The amount of feedback is also decreased significantly because the BS doesn't need to track each subchannel separately. Literature lacks QoS-based resource allocation algorithms for this scenario, therefore we especially focus our current work on the second class of subchannels, i.e., PUSC/FUSC.

Our proposed scheme first uses a variation of LWDF scheduling policy to find the rate requirements of real time users, which is explained in Section IV. Then, solving a constrained optimization problem formulated in Section V, power and bandwidth are allocated to users in a way to maximize proportional fairness for data users, while satisfying rate requirements for real time users. Frequency selective fading is also considered in Section VI and joint subchannel and power allocation methods are considered in order to improve the performance of existing schemes.

II. SYSTEM MODEL

We consider a multicarrier scheme where multiple access is provided by assigning a subset of subchannels to each receiver at each time frame. Let W and P denote the total bandwidth and power, respectively. Total bandwidth W is divided into K subchannels of bandwidth W_{sub} Hz, each consisting of a group of carriers.

The noise and interference power density is N_0, and the channel gain averaged over the entire band from the BS to user i at time t is $h_i(t)$, where $h_i(t)$ includes path loss, shadowing (log-normal fading) and fast fading. As explained in the Introduction section we assume distributed subcarrier grouping and therefore all subchannels are of equal quality with respect to a user.

There are three classes of users. Users in the classes U_D, U_S, and U_V demand data, video streaming, and voice traffic, respectively. Let D, S, and V be denote their quantities. Let $U_R = U_S \cup U_V$ be the set of users demanding real-time traffic. The system that we consider is time slotted with frame length T_s seconds. The scheduler makes a resource allocation decision at each time frame. We assume that the numbers of all types of sessions are fixed throughout the system simulation.

IEEE 802.16a/e standards allow several combinations of modulation and coding rates that can be used depending on the signal to noise ratio. Here assuming constant fading during a frame, we model the channel as an additive white gaussian noise (AWGN) channel. Base station allocates the available power and rate among users, where $p_i(t)$ and $w_i(t)$ are the power and bandwidth allocated to user i in frame t. For an SINR $\frac{p_i(t)h_i(t)}{N_0 w_i(t)}$, the highest order modulation and coding scheme that guarantees a BER constraint is used [21]. In this work, for simplicity we assume rate as a continuous function of SNR. The resulting SNR values can be further quantized to the available values, which is not considered in this work.

Based on the modulation/coding pairs and corresponding SNR thresholds in the standard, it is reasonable to approximate the optimal transmission rate as an increasing and concave function of the signal to noise ratio. We will adopt the Shannon channel capacity for AWGN channel as a function for bandwidth and

transmission power assigned to user i:

$$r_i(w_i(t), p_i(t)) = w_i(t) \log_2 \left(1 + \beta \frac{p_i(t) h_i(t)}{N_0 w_i(t)}\right). \quad (1)$$

The reason for using (1) is its simplicity, and it also approximates rate-SINR relation in the standard with $\beta = 0.25$. The parameter $0 < \beta < 1$ (SNR gap) compensates the rate gap between Shannon capacity and rate achieved by practical modulation and coding techniques [22].

III. BENCHMARK SCHEMES

As mentioned in the Introduction, previous works in [11], [12], [13], and [23] proposed resource allocation schemes for OFDMA-based systems supporting heterogeneous traffic. Below, we explain these algorithms. Let $h_{i,k}(t)$ be the channel gain of user i at subchannel k.

A. Largest Weighted Delay First with Proportional Fairness (LWDF-PF)

This is a resource allocation scheme that combines weighted delay based scheduling with proportional fairness [23]. Subchannel allocation is performed based on a utility value,

$$U_i(t) = \frac{\kappa_i D_i^{HOL}(t)}{R_i(t)} \log \left(1 + \frac{\beta P h_{i,k}(t)}{N_0 W}\right). \quad (2)$$

$D_i^{HOL}(t)$ is the head of line packet delay of user i. The parameter κ_i is a positive constant that reflects the QoS priority of the user/session. If QoS requirement is defined as delay violation probability, $P(D_i > D_i^{max}) < \delta_i$, where D_i^{max} is the delay constraint and δ_i is the probability of exceeding this constraint (typically 0.05), then the constant κ_i can be defined as $\kappa_i = -\frac{\log(\delta_i)}{D_i^{max}}$ [11], [23], and this weighted delay metric is used in channel allocation in single-channel time-sharing systems, and referred to as largest weighted delay-first (LWDF) scheme [10], [11], [23]. Here, $R_i(t)$ is the average received rate which is updated as,

$$R_i(t+1) = \alpha_i R_i(t) + (1 - \alpha_i) r_i(t) \quad (3)$$

where $0 < \alpha_i < 1$ is typically close to one. In the proportional fair scheme $T = 1/(1 - \alpha_i)$ is the length of the sliding time window and average rate is computed over this time window at each frame. This way we consider both current rate as well as rates given to the user in the past. Observed at time t, the highest consideration is given to the current rate $r(t)$, and the rates received at the past $t-1$, $t-2$, $\cdots$ carry diminishing importance. Including $R_i(t)$ provides proportional fairness. The LWDF-PF method can be summarized as follows:

1. Initialize $w_i = 0, r_i = 0, \Omega_i = \emptyset, \forall i$, where w_i, r_i, and Ω_i are the number subchannels, rate and set of subchannels for user i.
2. **for** k=1:K
   2-1. Select user

$$i^* = \arg \max_{i: q_i(t) > r_i T_s} U_i(t)$$

where $q_i(t)$ is the number of bits waiting to be transmitted to user i at time t.

2-2. Allocate subchannel k to user i^*, $\Omega_{i^*} = \Omega_{i^*} \cup \{k\}$.

2-3. Update received rate as

$$r_{i^*} \rightarrow \sum_{k \in \Omega_{i^*}} W_{sub} \log_2 \left(1 + \frac{\beta P h_{i^*,k}(t)}{N_0 W}\right)$$

2-4. Update queue size as

$$q_{i^*}(t) \rightarrow q_{i^*}(t) - W_{sub} \log_2 \left(1 + \frac{\beta P h_{i^*,k}(t)}{N_0 W}\right)$$

end

Step 2 avoids excessive allocation of resources to users and prevents underutilization. For flat fading channels, since $h_{i,k} = h_i, \forall k$ the received rate is updated as

$$r_{i^*} = |\Omega_i| W_{sub} \log_2 \left(1 + \frac{\beta P h_{i^*}(t)}{N_0 W}\right)$$

where $|\Omega_i|$ is the cardinality of the set Ω_i.

B. Channel-Aware Queue-Aware Scheduling with Joint Subchannel and Power Allocation (CAQA-JSPA)

This algorithm proposed in [13] extends LWDF in two ways: 1) The power allocated to each user is still proportional to the number subchannels, however a power given to a user can be optimally distributed to its subcarriers to maximize received rate with the given resources. Subchannels are still allocated one by one. Let w_i be the aggregate amount of bandwidth allocated to user i. The power allocated to this user is $P \cdot w_i/W$. This time received rate r_i is calculated by allocating this amount of power *optimally* to the w_i/W_{sub} subcarriers. This power optimization is repeated as each new subchannel is allocated to a user. The received rate for user i at each step is therefore found by solving

$$r_i = \max_{p_{i,k}} \sum_{k \in \Omega_i} W_{sub} \log_2 \left(1 + \frac{\beta p_{i,k} h_{i,k}(t)}{N_0 W}\right)$$
$$\text{s.t.} \sum_{k \in \Omega_i} p_{i,k} \leq P \cdot w_i / W.$$

This is solved by waterfilling. For the case of flat fading or distributed subcarrier grouping the optimal allocation is the uniform allocation, therefore this extension does not provide any improvement for this case. 2) As each subchannel is allocated and received rates are updated, head of line packet delay $\widehat{D}_i^{HOL}(t)$ is estimated based on the current received rate and arrival times of the unserved packets [13]. The following metric is used at step k,

$$U_{i,k}(t) = \frac{\kappa_i \widehat{D}_i^{HOL}(t)}{R_{i,k}(t)} \log \left(1 + \frac{\beta P h_{i,k}(t)}{N_0 W}\right)$$

where $R_{i,k}(t) = \alpha_i R_{i,k}(t-1) + (1 - \alpha_i) W_{sub} \log_2(1 + \frac{\beta p_{i,k}^*(t) h_{i,k}(t)}{N_0 W})$ is the average service that user i received from subchannel k. Here, $p_{i,k}^*(t)$ is the power allocated to user i and subchannel k at time t. For flat fading channels using separate

average rates for subchannels also don't make any difference because each subchannel is equal and subchannels are allocated randomly once the number of subchannels to be allocated to each user is determined.

The max delay utility (MDU) algorithm proposed in [11] uses mean queue size instead of head of line delay. Simulation results show that CAQA algorithm [12] achieves better performance than MDU, therefore it is not taken as a benchmark in this work.

The algorithms explained above are proposed for systems with random packet arrivals and delay constraints. Best effort traffic such as Web browsing or FTP, on the other hand have very loose or no delay constraints. Besides, we assume that data traffic source adjusts its transmission rate to suite the service rate and it can always use any bandwidth assigned to it. We can assume that transmission queue of data traffic is never empty. Under such assumptions there is no delay for data traffic. In order to be able to use metric (2) for data traffic we will assume a fixed (e.g., one second) delay for data sessions.

IV. PROPOSED SCHEME

Our primary aim is to find a scheduling scheme that supports data traffic as well as delay sensitive real-time traffic. Our approach is based on the assumption of frequency-flat fading. This implies that each subchannel is equal with respect to a user. Based on this assumption we consider total bandwidth as a continuously divisible quantity. Our solution for resource allocation consists of,

- Determining the rate constraints for users demanding real time traffic.
- Formulating and solving an optimization problem that aims proportional fairness for data users subject to rate constraint for real time users and total power/bandwidth constraints.
- Quantizing the resultant bandwidth to integer multiples of subchannel bandwidth.

We will first formulate the proportional fairness objective for data users.

A. Proportional Fairness for Data Traffic

It is proven in [4] by Tse that a proportional fair allocation for a single carrier system also maximizes the sum of the logarithms of average user rates $\sum_{i=1}^{N} \log R_i$ where N is the number of users and R_i is the average received rate of user i. In a single carrier system proportional fairness is achieved by scheduling at each frame t, a user i^* according to:

$$i^* = \arg\max_i \frac{r_i(t)}{R_i(t)}. \tag{4}$$

Here $r_i(t)$ is the instantaneous transmittable rate to user i at the current frame. $R_i(t)$ is the average data rate that user i receives over time. At each frame the average rate is updated according to (3). So this method maintains fairness in the long run, while trying to schedule the user with the best channel at each frame.

Proportional fair resource allocation problem in OFDMA systems was modeled previously in [6] and [7] as maximizing the sum of logarithms of instantaneous rates. In [7], it was studied for flat fading multichannel systems and formulated as a joint power and bandwidth optimization problem as $\sum_{i=1}^{N} \log\left(r_i(w_i, p_i)\right)$ subject to power constraint $\sum_{i=1}^{N} p_i \leq P$, bandwidth constraint $\sum_{i=1}^{N} w_i \leq W$ and $p_i, w_i \geq 0, \forall i$, where $r_i(w_i, p_i)$ is the *instantaneous* rate function in (1). In [7], efficient and low complexity algorithms are proposed to solve the above optimization problem. Some algorithms were also proposed for the same objective in [6] and [9]. However this objective function aims proportional fairness only in a single time slot as opposed to long term. In fact, data users do not need fairness in a very short term. Fair allocation in a few seconds (e.g., thousand frames) of time window is enough, since it takes that much time to download a webpage. This gives a degree of freedom and facilitates the use of time diversity in order to maximize the data rate and proportional fairness and it should be exploited.

Long term proportional fairness in multichannel systems was formulated in [5] and a solution was proposed in the recent work [8]. Here the authors propose a subchannel and time slot allocation scheme in order to maximize the sum of logarithms of long term received rates. They use the following greedy approach. At each time frame the long term average received rates up to the beginning of the current slot are given. The average rates are computed using a moving average formula similar to (3). The problem at each frame is to determine the current rates that maximize the log-sum of moving averages. In this work we follow a similar approach and consider the maximization of the following objective function.

$$C(\mathbf{r}(t)) = \sum_{i=1}^{N} \log\left(\alpha_i R_i(t-1) + (1-\alpha_i) r_i(w_i(t), p_i(t))\right).$$

After some rearrangements the objective for data users becomes:

$$\max_{\mathbf{p}(t),\mathbf{w}(t)} \sum_{i=1}^{N} \log\left(\alpha_i + \frac{(1-\alpha_i) r_i(w_i(t), p_i(t))}{R_i(t-1)}\right)$$
$$= \max_{\mathbf{p}(t),\mathbf{w}(t)} \prod_{i}^{N} \left(\alpha_i + \frac{(1-\alpha_i) r_i(w_i(t), p_i(t))}{R_i(t-1)}\right). \tag{5}$$

The novelty of our proposed scheme is that we also consider power optimization in addition to subchannel. Besides we include rate constraints for the voice and video streaming users as a constraint, which will be explained below.

B. Real Time Traffic

Proportional fairness objective in (5) aims at providing fairness to data users. On the other hand, real time traffic has more strict delay and packet loss requirements, which can be translated into strict instantaneous rate requirements.

In this work we propose to use a variation of the LWDF-PF algorithm in computing the rate requirements for real time users. The only difference is that as new subchannels are allocated to user after Step 2-3 average received rate $R_{i^*}(t)$ is updated as $R_{i^*}(t+1) = \alpha_{i^*} R_{i^*}(t) + (1-\alpha_{i^*}) r_{i^*}(t)$. This avoids excessive allocation to a user. LWDF-type of algorithms have good performance, however they can be improved by using power control. Let $w_i, \forall i \in U_R$ be the amount of bandwidth given to each real

$$L(\mathbf{w},\mathbf{p},\lambda_p,\lambda_w,\lambda^r)=\prod_{i\in U_D}\left(\alpha_i+\frac{(1-\alpha_i)w_i\log_2\left(1+\frac{p_i}{n_iw_i}\right)}{R_i}\right)+\lambda_p\left(P-\sum_{i\in U_D\cup U'_R}p_i\right)+\lambda_w\left(W-\sum_{i\in U_D\cup U'_R}w_i\right)$$
$$+\sum_{i\in U'_R}\lambda_i^r\left(w_i\log_2\left(1+\frac{p_i}{n_iw_i}\right)-r_i^c\right). \quad (12)$$

time session as a result of applying the scheme described above. Then the rate constraints for each real time session is calculated as

$$r_i^c=w_i\log_2\left(1+\frac{\beta Ph_i(t)}{N_0W}\right),\ \forall i\in U_R. \quad (6)$$

We will use r_i^c as the rate constraint in the constrained optimization problem to be defined in Section V. Solving that problem will help us achieve the same rate with less resources, which will leave more resources for the data users.

This way we propose a joint power/bandwidth optimization scheme that can be used on top of LWDF-type of uniform power schemes in order to improve the performance[1]. Delay constrained voice and video sessions have lower rate and they have to be served despite possibly bad channel conditions, therefore they are limited by power. On the other hand, delay tolerant data sessions are only scheduled when their SINR is good, therefore they are bandwidth limited. Our joint power and bandwidth allocation framework tends to give more power to the former and more bandwidth to the latter types of sessions. By comparing it directly to LWDF and CAQA, it will be easy to observe the effects of joint power/bandwidth control on the performance.

V. JOINT DATA AND REAL TIME RESOURCE ALLOCATION

In this section we combine the proportional fair scheduling objective in (5) and real time user rate requirements defined in (6) and propose an adaptive power and bandwidth allocation (APBA) scheme.

We formulate a constrained optimization problem, where the objective function is (5) and the constraints are the total power/bandwidth constraints and the rate requirements for real time sessions. Let $n_i=\frac{N_0}{\beta h_i}$. The resulting optimization problem is as follows:[2] Find

$$(\mathbf{p}^*,\mathbf{w}^*)=\arg\max_{\mathbf{p},\mathbf{w}}\prod_{i\in U_D}\left(\alpha_i+\frac{(1-\alpha_i)w_i\log_2\left(1+\frac{p_i}{n_iw_i}\right)}{R_i}\right) \quad (7)$$

subject to

$$\sum_{i\in U_D\cup U'_R}p_i^*\le P \quad (8)$$
$$\sum_{i\in U_D\cup U'_R}w_i^*\le W \quad (9)$$
$$w_i^*\log_2\left(1+\frac{p_i^*}{n_iw_i^*}\right)\ge r_i^c, i\in U'_R \quad (10)$$
$$p_i^*,w_i^*\ge 0,\forall i\in U_D\cup U'_R. \quad (11)$$

Here U'_R is the set of real time sessions with rate greater than zero.

A. Solution to the Constrained Optimization Problem

The objective function (7) is an increasing function of $(\mathbf{w},\mathbf{p})$, therefore the maximum is achieved only when the constraints (8), (9), and (10) are all met with equality. For this reason, we can replace these inequalities with equalities in the discussion below.

Lemma 1: The problem in (7), (8), (9), (10), and (11) is a convex optimization problem.

Proof: In the Appendix. □

Actually there is no guarantee that a solution can be found that satisfies (10) for all users. The rate requirements for real time users can be too high that it may be impossible to satisfy with the given channel conditions.

To start with, we assume that the problem is *feasible*. We will discuss about how to detect infeasibility of the problem and what to do in that case in the next section. We can write the Lagrangian of the problem as [24] in (12)

Taking the derivatives of $L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)$ with respect to p_i, w_i for all users, we get the following:

- For users $i\in U_D$:

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial p_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)}=0$$
$$\Rightarrow\lambda_p=\frac{1/(n_i\ln 2)}{\left(R_i\widetilde{\alpha}_i+w_i\log_2\left(1+\frac{p_i^*}{n_iw_i^*}\right)\right)\left(1+\frac{p_i^*}{n_iw_i^*}\right)} \quad (13)$$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial w_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)}=0$$
$$\Rightarrow\lambda_w=\frac{\left(1+\frac{p_i^*}{n_iw_i^*}\right)\log\left(1+\frac{p_i^*}{n_iw_i^*}\right)-\frac{p_i}{n_iw_i^*}}{\ln 2\left(1+\frac{p_i^*}{n_iw_i^*}\right)\left(R_i\widetilde{\alpha}_i+w_i^*\log_2\left(1+\frac{p_i^*}{n_iw_i^*}\right)\right)} \quad (14)$$

[1] Please note that the proposed solution is not exactly optimal. First of all we make a continuous bandwidth assumption as explained in Section IV. Secondly we determine rate constraints based on results of a LWDF-type uniform power subchannel allocation scheme, which may not be optimal. Besides the notion of optimality is unclear when there are contradictory objectives of maximizing proportional fairness for data users and providing QoS for real time sessions. However, our approach is based on solving a convex optimization problem, which is why we use the words *optimization*, or *optimal* throughout the paper.

[2] Here, p_i, w_i, and n_i are the values at time t. The time index is not shown for convenience.

where $\widetilde{\alpha}_i = \frac{\alpha_i}{1-\alpha_i}$. By dividing (14) with (13) we can write for all $i \in U_D$:

$$\frac{\lambda_w}{\lambda_p} = \Lambda_x = n_i \left((1+x_i^*)\log(1+x_i^*) - x_i^*\right) \quad (15)$$

where $x_i^* = \frac{p_i^*}{n_i w_i^*}$ denotes the optimal *effective* SINR, which is the SINR multiplied by the SINR gap parameter β.

- For users $i \in U_R'$:

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial p_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0$$
$$\Rightarrow \frac{\lambda_p}{\lambda_i^r} = \frac{1}{n_i \ln 2}\frac{1}{1+\frac{p_i^*}{n_i w_i^*}} \quad (16)$$
$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial w_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0$$
$$\Rightarrow \frac{\lambda_w}{\lambda_i^r} = \frac{1}{\ln 2}\left(\log\left(1+\frac{p_i^*}{n_i w_i^*}\right) - \frac{\frac{p_i^*}{n_i w_i^*}}{1+\frac{p_i^*}{n_i w_i^*}}\right). \quad (17)$$

Combining (16) and (17) (dividing $\frac{\lambda_w}{\lambda_p}$) for all $i \in U_R'$ again gives:

$$\frac{\lambda_w}{\lambda_p} = \Lambda_x = n_i \left((1+x_i^*)\log(1+x_i^*) - x_i^*\right). \quad (18)$$

By writing (18) we can eliminate λ_i^r's from the problem. It is worth noting that we get the same relation between Λ_x/n_i and x_i for all users ((15) and (18)). At this point, it is convenient to define the function $f_x(x)$ as:

$$f_x(x) = (1+x)\log(1+x) - x. \quad (19)$$

Then, we have

$$x_i(\Lambda_x) = f_x^{-1}(\Lambda_x/n_i), \forall i \in U_D \cup U_R'. \quad (20)$$

Lemma 2: Effective SINR ($x_i(\Lambda_x)$) is a monotonic increasing function of Λ_x for users $i \in U_D \cup U_R'$.

Proof: The proof is in Appendix. □

From (13) and (15) for data users we can write, for $i \in U_D$:

$$w_i(\lambda_p,\lambda_w) = \frac{\left[\frac{1}{\lambda_p} - \ln 2 n_i (1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))R_i\widetilde{\alpha}_i\right]^+}{\log(1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))(1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))n_i} \quad (21)$$

$$p_i(\lambda_p,\lambda_w) = \frac{\left[\frac{1}{\lambda_p} - \ln 2 n_i (1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))R_i\widetilde{\alpha}_i\right]^+ f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i})}{\log(1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))(1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))}. \quad (22)$$

The $[\cdot]^+ = \max(0,.)$ operator in (22) and (22) guarantees that $w_i, p_i \geq 0$ for all users.

- For users $i \in U_R'$ we have:

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial \lambda_i^r}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0$$
$$\Rightarrow r_i^c = w_i^* \log\left(1+\frac{p_i^*}{n_i w_i^*}\right), \forall i \in U_R'. \quad (23)$$

Using (18) and (23), we find,

$$w_i(\lambda_p,\lambda_w) = \frac{r_i^c}{\log(1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))} \quad (24)$$

$$p_i(\lambda_p,\lambda_w) = \frac{r_i^c f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}) n_i}{\log(1+f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i}))}. \quad (25)$$

Total power and bandwidth constraint equations are

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial \lambda_p}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0$$
$$\Rightarrow P = \sum_{i\in U_D\cup U_R'} p_i(\lambda_p^*,\lambda_w^*) \quad (26)$$
$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial \lambda_w}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0$$
$$\Rightarrow W = \sum_{i\in U_D\cup U_R'} w_i(\lambda_p^*,\lambda_w^*). \quad (27)$$

Given λ_p and λ_w we can compute the power and bandwidth for all users using (22)–(25). We need to find the right λ_p^* and λ_w^* so that the power and bandwidth constraints are satisfied. Let $S_p(\lambda_p,\lambda_w)$ and $S_w(\lambda_p,\lambda_w)$ be the total bandwidth and total power functions. The problem is finding λ_w^* and λ_p^* such that

$$S_w(\lambda_p^*,\lambda_w^*) = \sum_{i\in U_D\cup U_R'} w_i(\lambda_p^*,\lambda_w^*) = W \quad (28)$$

$$S_p(\lambda_p^*,\lambda_w^*) = \sum_{i\in U_D\cup U_R'} p_i(\lambda_p^*,\lambda_w^*) = P. \quad (29)$$

In fact because of the convexity of the problem its optimal solution is unique and it is the λ_w^* and λ_p^* values that satisfy the total power and bandwidth constraints. Before finding those values we have to first find out if the problem is feasible. The procedure for finding it is explained in the Appendix. If the problem is not feasible then the BS chooses the user that consumes the most power (probably the worst channel condition) and decreases its rate to half ($r_i^c \rightarrow r_i^c/2$). If $r_i^c < r_i^0$ (r_i^0 is the bit arrival rate for the session) then r_i^c is set to zero. The BS doesn't directly decrease its rate to zero, because the user with bad channel condition probably needs urgent service (otherwise it wouldn't have been given resource in the initial subchannel allocation phase), therefore the BS still tries to give a chance to that users. This rate decreasing procedure is repeated until the problem is feasible.

Finding the two optimal Lagrange multipliers require a two dimensional search. Ellipsoid method [25] is a subgradient-based optimization method that is a generalization of binary search for multiple dimensions. It is especially useful in optimizing non-differentiable functions. Our problem includes such non-differentiable functions because of the existence of the $[\cdot]^+$ operator in power/bandwidth evaluations. Briefly, the ellipsoid method involves finding an initial ellipsoid that is guaranteed to include the optimal values. Then, at each step of the algorithm a subgradient is found and the volume of the ellipsoid is decreased until it is small enough. The subgradients that are used and the ellipsoid updates are explained in the Appendix. Below, we prove some properties of the optimal Lagrange multipliers

λ_p^*, λ_w^* and $\Lambda_x^* = \frac{\lambda_w^*}{\lambda_p^*}$ that will be useful in finding an initial ellipsoid.

Lemma 3: The following inequalities hold:

$$\Lambda_x^* \geq \Lambda_x^{min} = \min_{i \in U_D \cup U_R'} \left\{ n_i f_x \left(\frac{P}{n_i W} \right) \right\}$$

$$\Lambda_x^* \leq \Lambda_x^{max} = \max_{i \in U_D \cup U_R'} \left\{ n_i f_x \left(\frac{P}{n_i W} \right) \right\} \quad (30)$$

$$\lambda_p \leq \frac{1}{\min_{i \in U_D \cup U_R'} \left\{ n_i (1 + f_x^{-1}(\Lambda_x^{\min}/n_i)) R_i \widetilde{\alpha}_i \right\}} \quad (31)$$

$$\lambda_w \leq \lambda_p^{\max} \Lambda_x^{\max}. \quad (32)$$

Proof: Proof is in the Appendix. □

The Lemma above gives us upper bounds for λ_p^* and λ_w^* and they are useful in defining the initial ellipsoid. After we find λ_w^* and Λ_p^*, we compute the optimal bandwidth and power values for all nodes. We will refer this scheme as APBA. This scheme has complexity of $O(N)$, because at each step of the iteration SNR values has to be computed for given Lagrange multipliers. In fact, it is less complex than that because only few of real time users have nonzero rate requirements at frame. Note that the complexity is independent of the number of subchannels.

B. Bandwidth Allocation for Uniform Power

This is a simpler alternative to joint power-bandwidth allocation. The rate function for user i becomes $w_i \log_2 \left(1 + \frac{P}{n_i W}\right)$. We can modify the problem (7)–(11) by using this rate function and excluding the power constraint in (8). By using standard techniques the optimal bandwidth values can be found by solving these set of equations,

$$w_i(\lambda_w) = \left[\frac{1}{\lambda_w} - \frac{R_i \widetilde{\alpha}_i}{\log_2 \left(1 + \frac{P}{n_i W}\right)} \right]^+, \forall i \in U_D \quad (33)$$

$$w_i(\lambda_w) = \frac{r_i^c}{\log_2 \left(1 + \frac{P}{n_i W}\right)}, \forall i \in U_R' \quad (34)$$

$$W = \sum_{i \in U_D \cup U_R'} w_i(\lambda_w). \quad (35)$$

Since bandwidth is a monotonic function of λ_w, optimal bandwidths can be found by binary search. Powers can be calculated as $p_i^* = P \frac{w_i^*}{W}$, $\forall i \in U_D \cup U_R'$. We will refer to this scheme as adaptive bandwidth allocation (ABA).

C. SINR/Bandwidth Quantization and Reshuffling

In practice, bandwidth allocation is in terms of integer number of subchannels. Hence, we have to apply the following resource shuffling procedure

- Quantize the bandwidth values to the nearest number of subchannel. Quantize to one subchannel if it is less than that.
- If the total bandwidth is greater than W, then find the node that has the largest increase in bandwidth due to quantization and decrease its one subchannel. If there is no such node left, then find user with the highest bandwidth and decrease by one subchannel.
- If total bandwidth is smaller than W, then find the node with the highest decrease in bandwidth due to quantization and increase its subchannels by one. If there is no such node left, then find user with best channel condition and give it one more bandwidth.

VI. FREQUENCY SELECTIVE CHANNELS

Most of the proposed OFDMA resource allocation schemes in the literature consider frequency selective fading and propose subchannel-by-subchannel allocation schemes in order to exploit frequency and multiuser diversity. Optimal solutions in such a scenario with our objectives of fairness and QoS is prohibitively complex. However, power control can still be used to improve the performance of of LWDF-type of schemes. In this case we divide the problem into two stages as the allocation for real-time and data users. The scheme that we use is summarized as follows

- Determine user rates and number of subchannels allocated to each user by applying an LWDF-type scheme.
- Based on the above information allocate subchannels in order to satisfy above rates with minimum power.
- Calculate the powers for the real time users. Keep their allocations fixed and reallocate the residual subchannels and power to the data users.

For the first step we apply the classical LWDF-PF algorithm explained in [23], except that we update average received rates as $\alpha_i R_i + (1 - \alpha_i) r_i$ after allocating each subchannel. This avoids excessive allocation to users. After the first step we obtain set of subchannels Ω_i allocated to user i and total rate $r_i^c = \sum_{k \in \Omega_i} W_{sub} \log_2(1 + \frac{\beta P h_{i,k}}{N_0 W_{sub}})$ for all users, where $h_{i,k}$ is the channel gain for user i at subchannel k.

Given the number of subchannels ($w_i = |\Omega_i|$) and rate constraints r_i^c found in the first step, we will reallocate the subchannels to satisfy those rate constraints with minimum power. The authors in [20] propose a linear programming-based solution to determine the subchannels allocated to each user in order to minimize the total power given the rate constraints and number of subchannels for each user. Even this solution is of high complexity, therefore a greedy scheme called Vogel's method is used and almost the same performance is achieved with much lower complexity. Once the subchannels are determined, waterfilling can be used in order to satisfy rate constraints with minimum power. In the second step we apply the Vogel's method explained as follows [20]:

1. Set number of subchannels $w_i = |\Omega_i|$, bits per symbol $S_i = \frac{r_i^c}{W_{sub} w_i}$, $\forall i = 1, \cdots, N$. Then, initialize $\Omega = \{1, 2, \cdots, K\}$ and $\Omega_i = \emptyset, \forall i$.
2. Set subchannel costs as $C_{i,k} = \frac{N_0 W_{sub}}{\beta h_{i,k}} (2^{S_i} - 1), \forall i, k$. Sort subchannel costs for each user in ascending order, which is denoted by C' (Now $C'_{i,1}$ is the cost of the lowest cost subchannel for user i).
3. Calculate user penalties as $P_i = C'_{i,w_i+1} - C'_{i,1}, \forall i$. This reflects the opportunity cost of not making an immediate allocation to a user. If the current best available channel(s) are very good, then the penalty of not allocating them to that user is high.

Table 1. Simulation parameters.

Parameter	Value
Cell radius	1.5km
User distances	0.3,0.6,0.9,1.2,1.5 km
Total power (P)	20 W
Total bandwidth (W)	10 MHz
Frame length	1 msec
Voice traffic	CBR 32 kbps
Video traffic	802.16 - 128 kbps
FTP file	5 MB
AWGN p.s.d.(N_0)	-174 dBm/Hz
Pathloss exponent (γ)	3.5
$\psi_{DB} \sim N(\mu_{\psi_{dB}}, \sigma_{\psi_{dB}})$	N(0 dB, 8 dB)
Coherent time (fast/slow)	(5 msec/400 msec.)
Pathloss(dB, d in meters)	$-31.5 - 35 \log_{10} d + \psi_{dB}$

Table 2. OFDMA-related parameters.

Parameter	Value
Nominal channel bandwidth	$W = 10$ MHz
FFT size	$N_{FFT} = 1024.$
Number of used subcarriers	$N_{used} = 840.$
Sampling factor	$n_s = F_s/W = 8/7$
Sampling frequency	$F_s = \lfloor n \times W/8000 \rfloor \times 8000 = 11.424$ MHz
Subcarrier spacing	$\Delta f = F_s/N_{FFT} = 1.1156 \times 10^4$ Hz
Used bandwidth	$N_{used} \times \Delta f = 9.37125$ MHz
Useful symbol time	$T_b = 1/\Delta f = 89.638\ \mu$s
Guards period ratio	1/8
OFDM symbol time	$T_s = (1 + 1/8) \times T_b = 0.1008$ msec
Subchannelization mode	DL-PUSC
Tones per subchannel	24
Subchannel bandwidth	$W_{sub} = 24 \times \Delta f = 267.744$ KHz
Number of subchannels	$K = 30$

4. **Repeat** until $w_i = 0, \forall i$
   5. Find user $i^* = \arg\max_{i:w_i>0} P_i$
      6. Find subchannel $k^* = \arg\min_{k\in\Omega} C_{i^*,k}$. Update $\Omega_i^* \rightarrow \Omega_i^* \cup \{k^*\}$, $\Omega \rightarrow \Omega/\{k^*\}$, $w_i^* = w_i^* - 1$ and $C_{i,k^*} = \infty, \forall i$. Sort C again and update penalties as in Steps 2 and 3.
7. **end Repeat**

At each repetition of lines 4 to 7, this algorithm allocates a subchannel to a user, so it ends in K steps. The subchannels are allocated to users in a way that requires close-to-minimum power to satisfy the rates obtained in the first step.

In the third step, given the subchannel allocations we further optimize powers. First, optimize power for real time users by using waterfilling. Then, we reallocate residual subchannels (Ω') and power (P') among the data users. Using the proportional fairness objective in (7) is still not practical for the frequency selective case. Considering that $1 - \alpha_i$ is small, the approximation $\ln(1 + x) \simeq x$ for small x can be used to convert it to the following

$$\max_{\Omega_i, p_{i,k}} \sum_{i\in U_D} \frac{1}{R_i} \left(\sum_{k\in\Omega_i} W_{sub} \log_2 \left(1 + \frac{\beta h_{i,k} p_{i,k}}{N_0 W_{sub}} \right) \right) \quad (36)$$

$$\sum_{i\in U_D} \sum_{k\in\Omega_i} p_{i,k} \leq P' \quad (37)$$

$$\Omega_i \cap \Omega_j = \emptyset,\ \forall i, j \in U_D \quad (38)$$

$$\bigcup_{i\in U_D} \Omega_i \subset \Omega'. \quad (39)$$

This is a weighted sum rate maximization problem subject to residual total power constraint and subject to the subchannel allocation constraints (No two user can share the same subchannel and subchannels should be allocated only from Ω'). This is a nonconvex problem , however using Lagrange dual relaxation techniques near-optimal solution is found in [18]. The solution is especially optimal as number of subchannels increases.

VII. PERFORMANCE EVALUATION

For the numerical evaluations we divide the users to 5 classes according to the distances, 0.3, 0.6, 0.9, 1.2, 1.5 km. At each distance level one fifth of the total number of users exist. We use the parameters in Table 1.

Table 2 summarizes the OFDMA-related parameters used in this simulation and their derivations. Here FFT size means the number of samples in the Fast Fourier Transformation. Number of used subcarriers N_{used} is smaller than N_{FFT} because the outer carriers in a subchannel do not carry modulation data.

We will compare our algorithm with the benchmark LWDF-PF [23] and CAQA-JSPA [13] schemes. Delay exceeding probability is taken as $\delta_i = 0.05$ for all users. Delay constraint for voice and video users are 0.1 and 0.4 seconds, respectively. For LWDF-PF algorithm we assume that the delay constraint for data users is 1 second and buffer length is infinite. We assume a constant HOL delay of 1 second for the data sessions. Filter values are $\alpha_i = 0.999, 0.995, 0.98$ for data, streaming and voice sessions.

Performance criteria are as follows. We will observe the total throughput for all data users. For data users we will also observe total log-sum rate $C(t) = \sum_{i\in U_D} \log R_i$. For real time users we will measure the outage probability , which means the percentage of the transmitted packets that violate their delay constraint.

A. Flat Fading

A.1 Increasing Number of Real Time Users

Figs. 1 and 2 show the effects of increasing the number of real time sessions on outage and data throughput performance. We assume that half of the real time users are voice and video streaming users. In Fig. 1, we observe that the proposed APBA scheme achieves the best throughput performance. APBA achieves 25% performance improvement with respect to benchmark algorithms especially when the number of real time users is large. The reason is that the proposed algorithm provides more power and less bandwidth to users with worse channel conditions therefore avoids the waste of bandwidth. The performance of the uniform-power ABA scheme is almost identical to the benchmarks, which again proves that *power control* is the process that makes the difference. We also observe that the performances of the two benchmarks are almost identical, where CAQA-JSPA achieves slightly more throughput (Fig. 1) but more outage (Fig. 2). We also see the performance improvement in terms of log-sum of throughput. Since it involves logarithms, the performance differences can't be high in numbers. Even small differences are important in this case.

In Fig. 2, we see the outage performances. We can observe that proposed APBA algorithms achieves less outage in addition

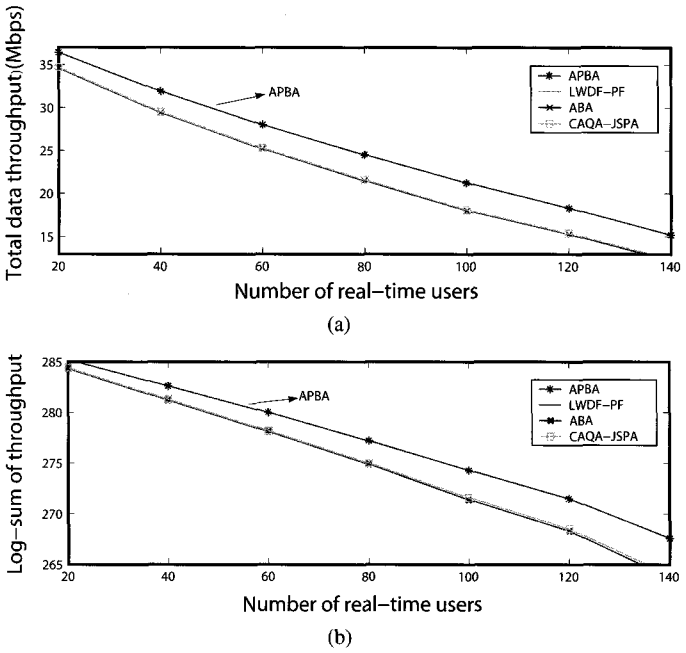


Fig. 1. Data user performance vs. number of real time users (D=20, S=V): (a) Total data throughput and (b) log-sum throughput for increasing number of real time users.

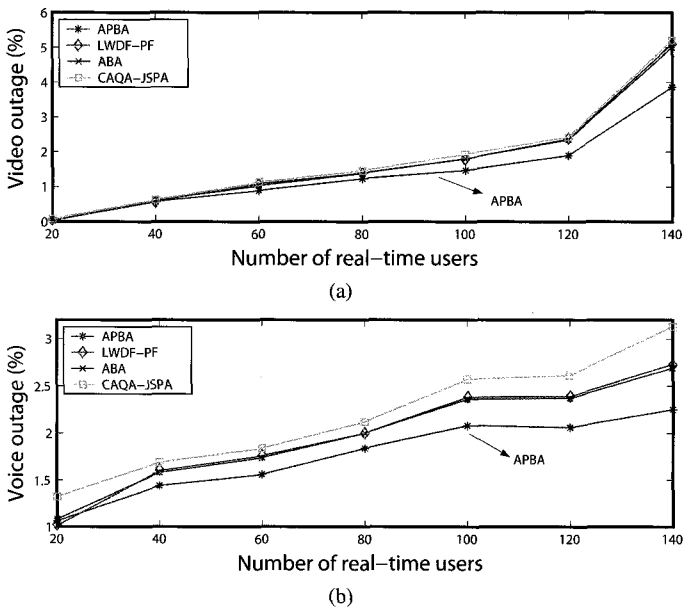


Fig. 2. Real-time user performance vs. number of real time user (D=20, S=V): (a) Outage probability for voice and (b) outage probability for video streaming sessions for increasing number of real time users.

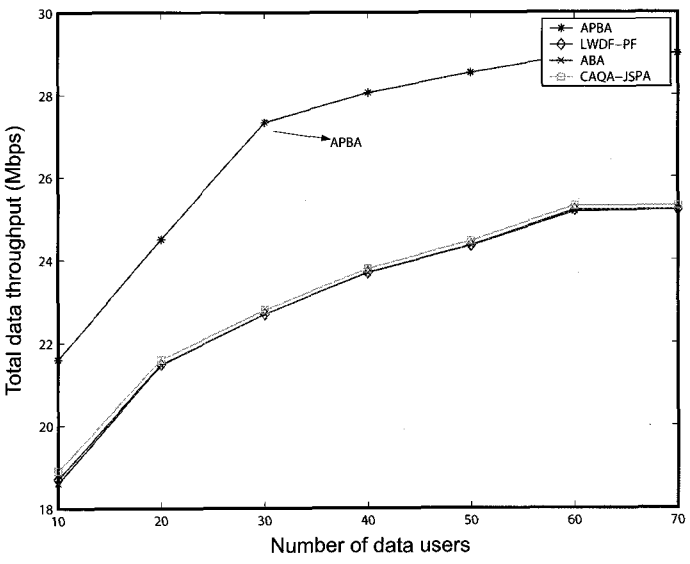


Fig. 3. Total data throughput for increasing number of data users (S=V=40).

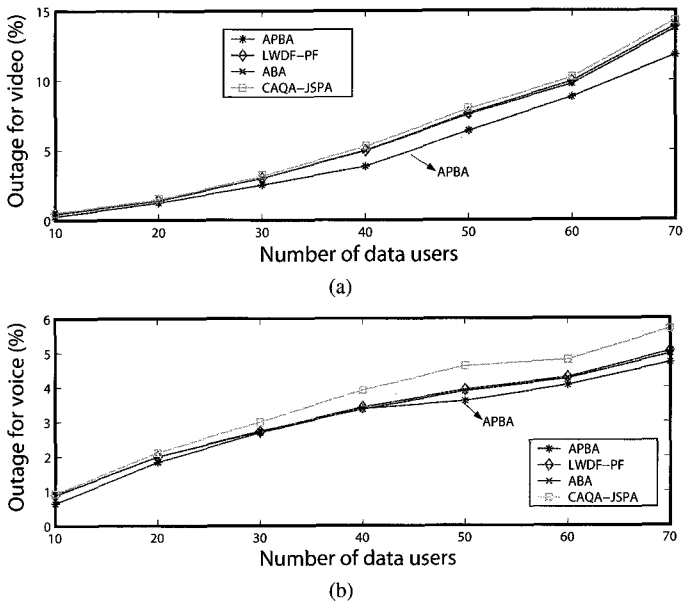


Fig. 4. Real time user performance vs. number of data users (S=V=40): (a) Outage for video and (b) Outage for voice for increasing number of data users.

to better throughput. Compared to the proposed scheme benchmarks have 25% higher outage for video when number of real time users is large. The improvement in voice sessions is even more.

A.2 Increasing Number of Data Users

In Figs. 3 and 4, we see the effects of increasing number of data users on data throughput and delay. From Fig. 3 it can be observed that data throughput is an increasing concave function of total number of data users. As number of users increase, multiuser diversity is in effect, however its effect diminishes as number of users increases. We also observe that there is a constant 15% performance difference between proposed APBA and benchmarks. We can conclude that proposed scheme provides more improvement when number of real time users is high. Although it is not included here, APBA has also better performance in terms of log-sum of throughput. Fig. 4 shows that this increase in throughput is obtained without sacrificing the outage performance.

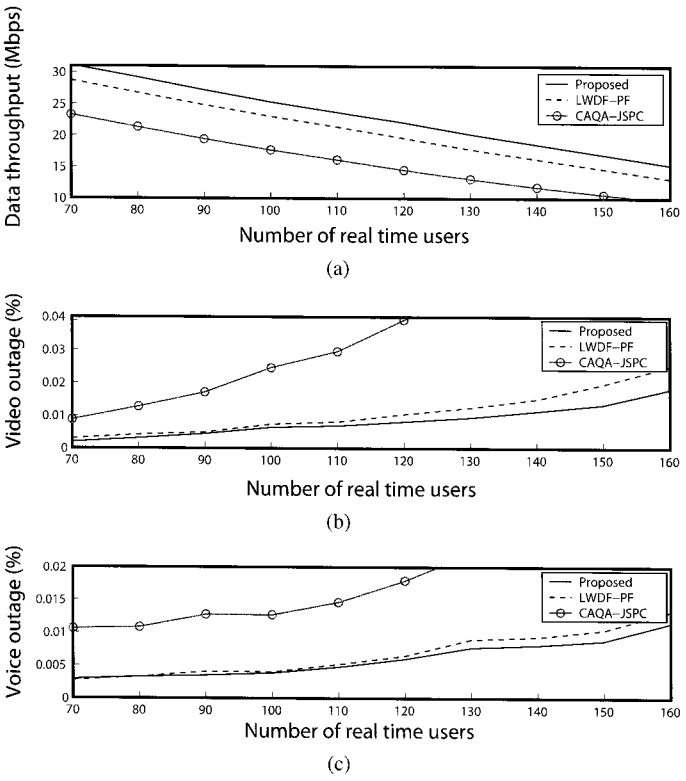


Fig. 5. Performance vs. number of real time users for the case of frequency selective fading (S=V, D=40): (a) Data throughput, (b) video outage, and (c) voice outage.

B. Frequency Selective Fading

Fig. 5 shows the performance for the case of frequency selective fading. First of all, here we observe that frequency selectivity especially improves the real time session outage performance significantly. We see that the proposed resource allocation scheme for frequency selective fading systems can provide $8-15\%$ (and more than 2 Mbps) performance improvement in terms of total data throughput. The percentage improvement increases as the number of real time users increase. The improvement with respect to LWDF-PF is less in frequency selective case than in the flat fading case. This is because exploiting the frequency diversity in frequency selective fading provides better SINRs for the users. The rate is a logarithmic function of SINR therefore power control provides diminishing improvement as SINR increases. We also observed that LWDF-PF performs better than CAQA-JSPA, although the latter also includes power control. This may be because, in CAQA-JSPA average received rate for each subchannel is considered separately. This does not reflect the actual rate that the user received. Besides power control in JSPA does not make a significant difference because a user still receives power proportional to number of channels it is allocated.

VIII. CONCLUSIONS

In this work we considered the problem of resource allocation in OFDMA based downlink wireless multiple access systems with flat fading or distributed subcarrier grouping. Such assumptions lead to simplified resource allocation policies in case of heterogeneous traffic requirements. We developed an optimization approach in order to provide proportional fairness for data users and satisfy delay requirements of real time users such as voice and video streaming. We solved the optimization problem and developed an algorithm that finds the optimal bandwidth and power given to each user. Our problem formulation also exploits time diversity for data users by considering the long term averaged received rate in optimization. The proposed scheme can be used on top of existing largest weighted delay first type of resource allocation policies and improve their performance by joint power/bandwidth control. Finally we numerically showed that when compared with the well known LWDF-PF and channel/queue aware (CAQA) schemes in the literature, our scheme both provides better proportional fairness for data sessions and provides better QoS for real time sessions.

IX. APPENDIX I

A. Proof of Concavity of Objective Function

The reward function

$$C(\mathbf{w}^n,\mathbf{p}^n)=\sum_{i\in U}\log\left(\alpha_i R_i+(1-\alpha_i)w_i\log_2\left(1+\frac{p_i}{n_i w_i}\right)\right) \tag{40}$$

is a concave function of w_i and p_i for all $i\in U_D$.

Proof: If we take the Hessian $\mathbf{H_r}$ of $r(p,w)=w\log_2(1+\frac{p}{nw})$

$$\mathbf{H_r}=\frac{-1}{(p+nw)^2}\begin{bmatrix} w & -p \\ p & \frac{p^2}{w}\end{bmatrix}, \tag{41}$$

we see that it is negative definite, therefore the function r is strictly concave. Multiplying it with a constant $(1-\alpha)$, adding it with a constant αR preserves concavity. Logarithm of a concave function is also concave. Linear combination (40) of concave functions are concave too; therefore reward function is a concave function of w_i and p_i for all $i\in U_D$.

□

B. Convexity of the Feasible Set

The feasible set of power and bandwidth levels (w,p) defined by (8), (9), (10), and (11) defines a convex set.

Proof: Consider two power-bandwidth vectors $(\mathbf{w}^1,\mathbf{p}^1)$ and $(\mathbf{w}^2,\mathbf{p}^2)$ that are in the feasible set. Now let us consider power-bandwidth vector $(\lambda\mathbf{w}^1+(1-\lambda)\mathbf{w}^2,\lambda\mathbf{p}^1+(1-\lambda)\mathbf{p}^2)$. It is clear that this vector satisfies the feasibility constraints in (11).

Now consider a user $i\in U'_R$. This user has a rate constraint r_i^c in (10). If (w_i^1,p_i^1) and (w_i^2,p_i^2) both satisfy constraint (10):

$$r(w_i^1,p_i^1)=w_i^1\log_2(1+\frac{p_i^1}{n_i w_i^1})=r_i^c \tag{42}$$

$$r(w_i^2,p_i^2)=w_i^2\log_2(1+\frac{p_i^2}{n_i w_i^2})=r_i^c,\forall i\in U'_R \tag{43}$$

From the concavity of the Shannon capacity with respect to w_i and p_i, we can write ($\overline{\lambda} = 1 - \lambda$):

$$r(\lambda w_i^1 + \overline{\lambda} w_i^2, \lambda p_i^1 + \overline{\lambda} p_i^2)$$
$$= (\lambda w_i^1 + \overline{\lambda} w_i^2) \log_2 \left(1 + \frac{\lambda p_i^1 + \overline{\lambda} p_i^2}{n_i(\lambda w_i^1 + \overline{\lambda} w_i^2)}\right) \quad (44)$$
$$r(\lambda w_i^1 + \overline{\lambda} w_i^2, \lambda p_i^1 + \overline{\lambda} p_i^2)$$
$$\geq \lambda w_i^1 \log_2 \left(1 + \frac{p_i^1}{n_i w_i^1}\right) + \overline{\lambda} w_i^2 \log_2 \left(1 + \frac{p_i^2}{n_i w_i^2}\right) \quad (45)$$
$$r(\lambda w_i^1 + \overline{\lambda} w_i^2, \lambda p_i^1 + \overline{\lambda} p_i^2) \geq \lambda r_i^c + \overline{\lambda} r_i^c = r_i^c. \quad (46)$$

Hence the power bandwidth values $(\lambda w_i^1 + \overline{\lambda} w_i^2, \lambda p_i^1 + \overline{\lambda} p_i^2)$ also satisfy the rate constraints for users $i \in U_R'$. □

C. Proof of Lemma 2

Proof: Effective SINR ($x_i(\Lambda_x)$) is a monotonic increasing function of Λ_x for users $i \in U_D \cup U_R'$.

Taking the derivative $\frac{df_x(x)}{dx} = \log(1 + x) > 0$, therefore $f_x(x)$ is an increasing function of x. Therefore its inverse $f_x^{-1}(\Lambda_x/n_i)$ is also an increasing function of Λ_x. For a given values of Λ_x/n_i, $f_x^{-1}(\Lambda_x/n_i)$ can be found by applying Newton's method (typically in as few as three iterations with 0.1% accuracy.)

□

D. Feasibility of the Solution

In the previous section we stated that there exists a solution to the problem, if the rate constraints are feasible. We now consider how to detect an infeasible problem and what to do in that case. If the problem is feasible (i.e. if the available power and bandwidth is enough to satisfy rate requirements of real time sessions), then there exists a $\Lambda_x = n_i f_x(x_i), \forall i \in U_R'$ so that the following inequalities hold:

$$\sum_{i \in U_R'} \frac{r_i^0}{\log\left(1 + f_x^{-1}\left(\frac{\Lambda_x}{n_i}\right)\right)} \leq W, \quad (47)$$

$$\sum_{i \in U_R'} \frac{r_i^0 f_x^{-1}\left(\frac{\Lambda_x}{n_i}\right) n_i}{\log\left(1 + f_x^{-1}\left(\frac{\Lambda_x}{n_i}\right)\right)} \leq P. \quad (48)$$

We know for all users $i \in U_D \cup U_R'$ that $x_i = f_a^{-1}(\Lambda_x/n_i)$ is increasing in Λ_x. The expression $\frac{x_i}{\log(1+x_i)}$ is strictly increasing in x_i. From these properties we can deduce that the left hand side (LHS) of (47) is a decreasing and LHS of (48) is an increasing function of Λ_x. Let Λ_x^0 be the smallest Λ_x value that satisfies the inequality 47. There exists such a $\Lambda_x^0 > 0$ because LHS of (47) is a strictly decreasing function which is infinity for $\Lambda_x = 0$ and zero for $\Lambda_x = \infty$. The problem is feasible *if and only if* RHS of (48) is smaller than P.

- $\Rightarrow$: If the LHS of (48) is smaller than P then both feasibility conditions (47) and (48) hold, therefore the problem is feasible.
- $\Leftarrow$: If the problem is feasible, then there exists Λ_x such that both (47) and (48) hold. Now let's assume that LHS of (48) is greater than P, then because of the monotonicity of the functions it is also greater for all $\Lambda_x \geq \Lambda_x^0$. Note that the LHS of (47) is greater than W for $\Lambda_x < \Lambda_x^0$. This means that there is no Λ_x such that both (47) and (48) hold and the problem is infeasible. This is a contradiction, therefore the LHS of (48) is smaller than P. The property holds.

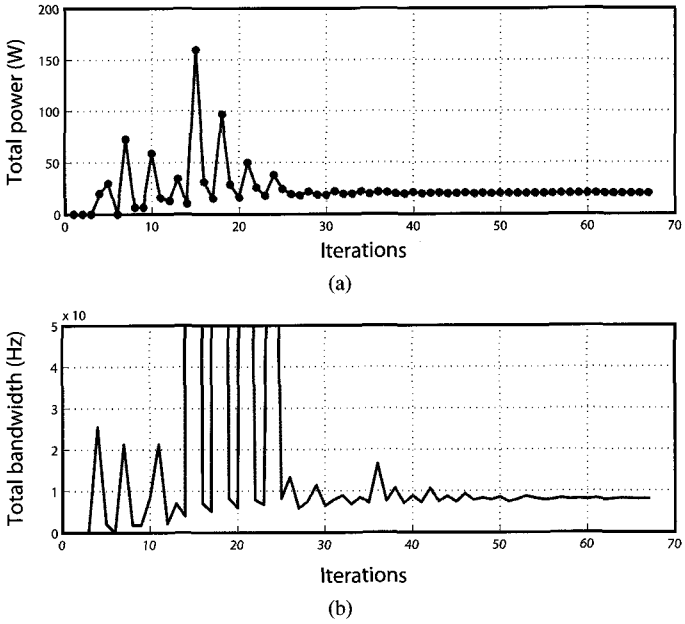


Fig. 6. Convergence of the total power and bandwidth using Ellipsoid Method: (a) Total power and (b) total bandwidth.

E. Proof of Lemma 3

- We can prove the inequalities for the optimal Λ_x^* using contraidction. Suppose that $\Lambda_x^* > \max_{i \in U_D \cup U_R'} \{n_i f_x(P/n_i W)\}$, $\forall i \in U_D \cup U_R'$, then $f_x^{-1}(\Lambda_x^*/n_i) > \frac{P}{n_i W}, \forall i$, from the monotonicity property. Then, the total power is greater than $\sum_{i \in U_D' \cup U_R'} w_i^* \frac{P}{n_i W} = P$, which contradicts with the power constraint, therefore the upper bound is proven. For the lower bound assume that $\Lambda_x^* < \min_{i \in U_D \cup U_R'} \{n_i f_x(P/n_i W)\}, \forall i \in U_D \cup U_R'$, then $f_x^{-1}(\Lambda_x^*/n_i) < \frac{P}{n_i W}, \forall i$, from the monotonicity property. Then, the total power is smaller than $\sum_{i \in U_D' \cup U_R'} w_i^* \frac{P}{n_i W} = P$. This is not optimal because proportional fairness metric can be increased by using the residual power, therefore the lower bound is also proven.
- For a feasible problem, bandwidth w_i of at least one data user should be greater than zero. From (22) we can write this as $\frac{1}{\lambda_p} \geq \min_{i \in U_D}\{n_i(1 + f_x^{-1}(\Lambda_x/n_i))\}$. By writing $\Lambda_x^{\min}$ instead of Λ_x and rearranging, we can write (32).
- From $\Lambda_x = \frac{\lambda_w}{\lambda_p}$ we can easily write (32).

F. Implementation of the Ellipsoid Method

The optimal $\Lambda = [\Lambda_w \lambda_p]^T$ is guaranteed to be contained in $\Lambda^{max} = [\Lambda_w^{max} \lambda_p^{max}]^T$. The initial value is chosen as $\Lambda_0 = 0.5 \times \Lambda^{max}$. The square matrix P_0 is defined as $A = 0.5 \times \text{diag}([\lambda^{max}]^2)$. Then, $(\lambda - \lambda_0)^T A^{0^{-1}} (\lambda - \lambda_0) \leq 1$ forms the initial ellipsoid.

At each step , subgradient d is chosen.

- If $\lambda_w^n < 0$ or $\lambda_p^n < 0$ then $d = -[-\lambda^n]^+$.
- If $\lambda_w^n, \lambda_p^n > 0$ then $d = [W - \sum_i w_i(\lambda_w^n, \lambda_p^n)P - \sum_i p_i(\lambda_w^n, \lambda_p^n)]^T$.

Updates are as follows [25]:

- $\widetilde{g} = d/\sqrt{d^T A d}$
- $\lambda^{n+1} = \lambda^n - \frac{1}{3}A^n\widetilde{g}$
- $A^{n+1} = \frac{4}{3}\left(A^n - \frac{2}{3}A^n\widetilde{g}\widetilde{g}^T A^n\right)$

Fig. 6 illustrates the characteristics of the sum-powers $S_p(\lambda_w, \lambda_p)$ and sum-bandwidth $S_w(\lambda_w, \lambda_p)$ for 20 data users and 20 real time users at one point in time. From the graph we see that indeed sum-power and sum-bandwidth converges to the constraint values. Because of the convexity of the problem convergence always occurs and the obtained Lagrange multipliers give the optimal power and bandwidth values.

REFERENCES

[1] C. Eklund, R. B. Marks, K.L. Stanwood, and S. Wang, "IEEE standard 802.16: A technical overview of the wireless MAN air interface for broadband wireless access," *IEEE Commun. Mag.*, June 2002.

[2] H. Ekstrom, A. Furuskar, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, and M. Wahlqvist, "Technical solutions for the 3G Long-Term Evolution," *IEEE Commun. Mag.*, pp. 38–45, Mar. 2006.

[3] R. Knopp and P.A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC*, 1995.

[4] D. Tse, "Forward link multiuser diversity through rate adaptation and scheduling." *submitted to IEEE J. Sel. Areas Commun.*, 2001.

[5] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.*, pp. 210–212, Mar. 2005.

[6] G. Song and G. Li, "Utility-based resource allocation and schedulng in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, Dec. 2005.

[7] C. Zhu and J. Agre, "Proportional-fair scheduling algorithms for OFDMA-based wireless systems," *Preprint, Fujitsu Labs*, 2006.

[8] M. Kaneko, P. Popovski, and J. Dahl, "Proportional fairness in multicarrier system with multislot frames: Upper bound and user multiplexing algorithms," *IEEE Trans. Wireless Commun.*, pp. 22–26, Jan. 2008.

[9] T. Nguyen and Y. Han, "A proportional fairness algorithm with QoS provision in downlink OFDMA systems" *IEEE Commun. Lett.*, pp. 760–762, Nov. 2006.

[10] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, pp. 150–154, Feb. 2001.

[11] G. Song, Y. Y. Li, L. J. Cimini, and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc. IEEE WCNC*, Mar. 2004, pp. 1939–1944.

[12] P. Parag, S. Bhashyam, and R. Aravind, "A subcarrier allocation algorithm for OFDMA using buffer and channel state information," in *Proc. IEEE VTC*, Sept. 2005, pp. 622–625.

[13] C. Mohanram and S. Bhashyam, "Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM," *IEEE Trans. Wireless Commun.*, pp. 3208–3213, Sept. 2007.

[14] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser subcarrier allocation for OFDM transmission using adaptive modulation," in *Proc. IEEE VTC*, May 1999, pp. 479–483.

[15] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. IEEE VTC*, May 2000, pp. 1085–1089.

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

[17] H. Kim, Y. Han, and S. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems," *IEEE Commun. Lett.*, pp. 526–528, June 2005.

[18] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. IEEE ISIT*, July 2006, pp. 1394–1398.

[19] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for ofdm systems," in *Proc. 40th Annual Conference on Information Sciences and Systems*, Mar. 2006, pp. 1272–1279.

[20] I. Kim, I.-S. Park, and Y. H. Lee, "Use of linear programming for dynamic subcarrier and bit allocation in multiuser OFDM," *IEEE Trans. Veh. Technol.*, pp. 1195–1207, July 2006.

[21] IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systesm, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobiel Operation in Licensed Bands and Corrigendum 1, *IEEE*, Feb. 2006.

[22] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, pp. 884–895, June 1999.

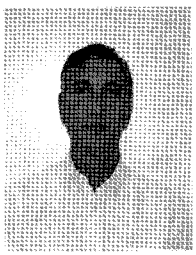
[23] G. Song, "Cross-layer resource allocation and scheduling in wireless multicarrier networks," Ph.D. Dissertation, Georgia Institute of Technology, Apr. 2005.

[24] L. Vanderberghe and S. Boyd, *Convex Optimization*. Mar. 2004.

[25] S. Boyd, *Ellipsoid Method*. Stanford University Class Notes, [Online]. Availabe: http://www.stanford.edu/class/ee392o/elp.pdf

Tolga Girici received his B.S. degree from Middle East Technical University, Ankara Turkey in 2000 and Ph.D. degree from University of Maryland, College Park in 2007, both in Electrical Engineering. He has been a research assistant at Institute of Systems Research from 2000 to 2005. In 2005 he has spent six months as an intern at the Intelligent Automation Inc, Rockville, MD, USA. He was a research assistant at the Fujitsu Labs, College Park MD, USA in 2006–2007. He is currently an assistant professor at TOBB University of Economics and Technology, Ankara, Turkey. His research interests include broadband cellular wireless access networks, satellite systems, and wireless ad hoc networks; network functions like multiple access, routing, broadcasting, resource allocation; optimization of performance objectives like fairness, energy efficiency, and quality of service.

Chenxi Zhu received his B.S. from Tsinghua University in Beijing, China in 1993 and Ph.D. from the University of Maryland, College Park in 2001. From 2001 to 2004 he was with Flarion Technologies Inc. in Bedminster, New Jersey working on flash-OFDM, an OFDMA-based mobile broadband wireless access network. Since November 2004 he has been a research scientist at Fujitsu Laboratories of America, where he has worked on wireless-LAN, WiMAX, and LTE.



Jonathan R. Agre is Vice President and General Manager at the Fujitsu Laboratories of America located in College Park, Maryland where he has helped to establish the Trusted Cybersystems Research Center. The center is focused on research to increase the trustworthiness of computer and communication systems. Through the center, he is also active in several standardization activities such as the WiMAX Forum, the Trusted Computing Group, and the INCITS M1 Biometrics group. Currently, he is working on security of ad hoc networks and standardization of vascular biometrics. He obtained the Ph.D. in Computer Science from the University of Maryland, in 1981 in performance modeling of distributed systems. He has over 70 technical articles and papers published in conferences, journals and books and over 35 patent applications. Prior to joining Fujitsu, he was at the Jet Propulsion Laboratory (JPL) working in advanced communication research and at the Rockwell International Science Center specializing in distributed systems for sensor networks and real-time factory control systems.

Anthony Ephremides holds the Cynthia Kim Professorship of Information Technology at the Electrical and Computer Engineering Department of the University of Maryland in College Park where he holds a joint appointment at the Institute for Systems Research, of which he was among the founding members in 1986. He obtained his Ph.D. in Electrical Engineering from Princeton University in 1971 and has been with the University of Maryland ever since. He has held various visiting positions at other Institutions (including MIT, UC Berkeley, ETH urich, INRIA, etc) and co-founded and co-directed a NASA-funded Center on Satellite and Hybrid Communication Networks in 1991. He has been the President of Pontos, Inc, since 1980 and has served as President of the IEEE Information Theory Society in 1987 and as a member of the IEEE Board of Directors in 1989 and 1990. He has been the General Chair and/or the Technical Program Chair of several technical conferences (including the IEEE Information Theory Symposium in 1991 and 2000, the IEEE Conference on Decision and Control in 1986, the ACM Mobihoc in 2003, and the IEEE Infocom in 1999). He has served on the Editorial Board of numerous journals and was the Founding Director of the Fairchild Scholars and Doctoral Fellows Program, a University-Industry Partnership from 1981 to 1985. He has received the IEEE Donald E. Fink Prize Paper Award in 1991 and the first ACM Achievement Award for Contributions to Wireless Networking in 1996, as well as the 2000 Fred W. Ellersick MILCOM Best Paper Award, the IEEE Third Millennium Medal, the 2000 Outstanding Systems Engineering Faculty Award from the Institute for Systems Research, and the Kirwan Faculty Research and Scholarship Prize from the University of Maryland in 2001, and a few other official recognitions of his work. He also received the 2006 Aaron Wyner Award for Exceptional Service and Leadership to the IEEE Information Theory Society. He is the author of several hundred papers, conference presentations, and patents, and his research interests lie in the areas of Communication Systems and Networks and all related disciplines, such as Information Theory, Control and Optimization, Satellite Systems, Queueing Models, Signal Processing, etc. He is especially interested in Wireless Networks and Energy Efficient Systems.