

Efficient Scheduling for Heterogeneous Services in OFDMA Downlink

Rajiv Agarwal[†], Vinay Majjigi[†], Rath Vannithamby[‡]
and John M. Cioffi[†]

[†]{rajivag, vmajjigi, cioffi}@stanford.edu [‡]rath.vannithamby@intel.com

Abstract—We consider optimal allocation of resources to users in a downlink OFDMA system to support heterogeneous applications consisting of both deadline-sensitive (DS) and best-effort (BE) data in a cross-layer manner. Given the user queue-states and long-term channel statistics, the proposed persistent scheduling algorithm allocates the minimum resources to ‘just’ meet the deadlines for DS packets (QoS guarantee). The remaining resources are optimally split between the users for their BE data to maximize sum-rate, minimize buffer holding cost or any general utility function. We formulate the resource allocation problem as a single geometric program (GP) that can be solved using standard convex optimization software tools. Simulation results are presented to compare the performance of different objectives for the BE applications in presence of DS traffic.

I. INTRODUCTION

Next generation wireless cellular networks will support a variety of quality-of-service (QoS)-sensitive applications like streaming multimedia and high-speed data for downlink users. To support such high-speed wireless services, transmission over a wide band using OFDMA is an attractive downlink transmission technique. To utilize system resources in an efficient manner, channel- and QoS-aware downlink transmission techniques play a key role. Efficient transmission schemes for the OFDMA downlink require optimal resource, namely: tone, rate and power allocation to the users based on their current channel states and backlogged packets.

Resource allocation for the OFDMA downlink has attracted much research attention. Under the framework of dynamic resource allocation (considered in this paper), resource allocation adapts to changes in users’ channel and queue conditions. For dynamic resource allocation, fast and less complex transmission schemes are particularly interesting, even if they may not result in a globally optimal solution.

Many scheduling schemes for the OFDMA downlink have been proposed, for example see [1] and references therein. In literature, resource allocation has been studied under two broad frameworks - F1) involves minimizing total transmit power for given rate constraints and F2) involves maximizing (weighted) sum-rate subject to a transmit power constraint. In either case, the optimal resource allocation, in general, is a difficult combinatorial problem [2]. The authors of [1], [2], [3], [4], [5], [6], [7] propose simpler approximations to the optimal problem.

This paper studies the problem of optimal resource allocation in the OFDMA downlink to support users with heterogeneous application requirements. Some users require deadline-sensitive (voice/video) and others may need best-

effort (data) service. It is assumed that the deadline-sensitive (DS) users have a higher priority than the best-effort (BE) users. The proposed algorithm computes the minimum system resources to ‘just’ meet the deadline-requirements of DS users. The remaining system resources (if any) are then optimally distributed between the BE users following one of the following scheduling disciplines - i) Queue Proportional Scheduling (QPS), ii) Maximum Weight Matching Scheduling (MWMS), iii) Best Channel Highest Possible Rate (BCHPR) or iv) Longest Queue Highest Possible Rate (LQHPR). The relative merits of these four have been studied in [8] and any one may be implemented in practice by a design engineer.

Dynamic resource allocation requires periodically informing the users of a change in allocation strategy. The time-duration between resource allocation updates is known as the scheduling period (SP). The authors of [1]-[6] consider a SP on the same order as the coherence time, resulting in a constant channel state over a SP. The implications of their assumption, especially with a short coherence-time, is the excessive overhead of informing users of a change in tone, transmit power and rate allocation. Similar to the approach in [7], in this paper, resource allocation is done by taking into account the long-term statistics of the channel. Since the long-term statistics change on a much larger time-scale, the SP duration can be kept large, much longer than the coherence-time of the channel. The proposed resource allocation algorithm, thus, performs optimal resource allocation over several channel realizations (arising from an ergodic fading distribution) in a single-run. This is attractive from the point of view of less complexity and less overhead of informing the users of allocation changes. Since optimal tone allocation is a hard combinatorial problem, the complexity of resource allocation and the overhead of informing users are further reduced by keeping the tone allocation to the users fixed for the entire duration of the SP. This is also known as persistent scheduling [9].

The work in this paper has several differences from those studied in [1]-[7]. Unlike schemes in [1]-[7], where a single framework (either F1 or F2) is used for all users based on their QoS requirements; this paper considers heterogeneous traffic and combines F1 and F2 for DS and BE users, respectively. The resource-allocation solution considers both the long-term and instantaneous channel knowledge; and tone allocation depends on long-term channel statistics only. The optimization problem is formulated as a Geometric Program (GP), which is a standard convex optimization problem that can be solved by

standard software packages like Matlab, Yalmip, cvx, etc. A similar problem was solved for the single-carrier case in [10]. Numerical results are provided to compare the performance of QPS with BCHPR (same as sum-rate maximization) for the BE users, in the presence of DS users when system parameters like number of tones, channel gain are varied.

Notation: $x \in \mathcal{C}$ denotes x is a complex number. $|\mathcal{S}|$ denotes the cardinality (the number of elements) of a set \mathcal{S} . $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ denotes set-union i.e. \mathcal{S} contains all elements of \mathcal{S}_1 and \mathcal{S}_2 . $\mathcal{S} = \mathcal{S}_1 \setminus \mathcal{S}_2$ denotes set-difference i.e. \mathcal{S} contains all elements of \mathcal{S}_1 which are not present in \mathcal{S}_2 . $x := y$ denotes the value of y is assigned to x , to distinguish from $x = y$, which states a mathematical fact.

II. SYSTEM MODEL

Consider an OFDMA Broadcast Channel with K users and M tones. Each tone undergoes independent and identically distributed (i.i.d.) block fading according to a stationary distribution. In a block, if tone m is assigned to user k at a transmission instant n , the received signal $y_{km}[n]$ is expressed as

$$y_{km}[n] = h_{km}x_{km}[n] + z_{km}[n], \quad k = 1 \text{ to } K, m = 1 \text{ to } M, n = 1 \text{ to } N_S \quad (1)$$

where $x_{km}[n]$ is the transmitted symbol, h_{km} is the complex channel-gain, and $z_{km}[n]$ is the i.i.d. zero-mean complex Gaussian noise with unitary variance. N_S is the number of symbols within a block for which h_{km} is constant¹. Both the transmitter and the receivers have perfect knowledge of h_{km} , which is reasonable for slow block fading assumed here.

The fading distribution of h_{km} is discrete. Γ_{km} denotes the set consisting of the possible values of channel power-gain γ_{km} defined as $\gamma_{km} \triangleq |h_{km}|^2$. The channel power-gain γ_{km} takes the i th value, γ_{km}^i , in the set Γ_{km} with probability p_{km}^i , $\sum_{i=1}^{|\Gamma_{km}|} p_{km}^i = 1$.

The base station allocates resources every scheduling period (SP) which has duration T_s seconds. T_s is much larger than the coherence time (Δt_c) of the channel, such that many channel realizations are seen every SP. Let N_B denote the number of blocks transmitted per SP, then it follows that N_B is large. This assumption begets the use of ergodicity for the fading process i.e. during the transmission of N_B blocks, $p_{km}^i N_B$ blocks see the channel in state i , $\forall k, m, i$.

Consider two classes of users: Deadline-Sensitive (DS) and Best-Effort (BE) users. Of the K users, K^{ds} are DS and the remaining K^{be} are BE users ($K^{ds} + K^{be} = K$). \mathcal{K}^{ds} and \mathcal{K}^{be} denote the set of DS and BE users, respectively. There is a deadline for each packet of a DS user that must be met. If a packet deadline expires, the packet is dropped. There are no deadlines for the packets of BE users.

For a DS user, the deadline associated with the application D_k , the packet-arrival process, and the QoS guarantee q_k imply a minimum average required-rate R_k^{req} . Typical values of q_k can be 0.9 or 0.95. Since, packet-delay is a function of

¹ $N_S = \frac{\Delta t_c}{T_{sym}}$, where T_{sym} is the symbol-duration.

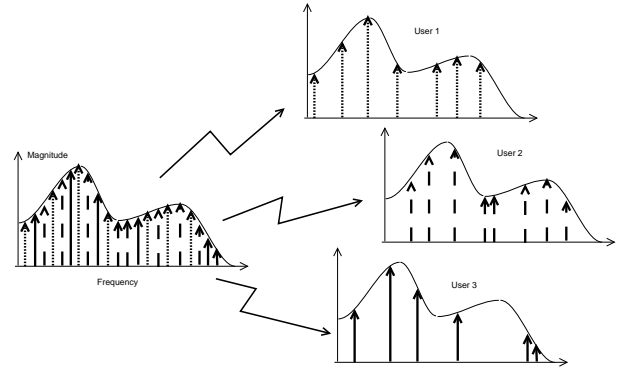


Fig. 1. OFDMA Broadcast Channel [7]

the packet arrival-process and the average transmission rate R_k , mathematically, the following relationship must hold for guaranteeing QoS to user k :

$$\text{Prob}(\text{Packet-delay}(\text{Arrival-Process}_k, R_k) > D_k) < 1 - q_k \quad (2)$$

Each user, given its packet-arrival process distribution and application deadline, calculates R_k for a given q_k using Eqn (2). R_k is the average transmission rate and its distribution follows the distribution of the channel gain. BE users do not have any deadline for their packets. Hence, their rate allocation is flexible and any general BE objective may be used to assign rates to the BE users. In Section IV, we compare the performance of two such BE objectives - QPS and BCHPR.

III. PROPOSED SCHEDULING SCHEME

This section describes the proposed scheduling policy and a coalition-based implementation of the algorithm based on [11]. If user k is allocated tone m , the power consumption P_{km}^i when the channel is in the i th fading state is

$$P_{km}^i = \frac{e^{s_{km}^i} - 1}{\gamma_{km}^i}, \quad (3)$$

which is obtained by inverting the capacity expression²

$$s_{km}^i = \log(1 + \gamma_{km}^i P_{km}^i) \quad (4)$$

A. Geometric Program Formulation

The optimization problem is to maximize a general rate-objective function for the BE users while satisfying minimum rate-requirements for DS users, an average power constraint³, and the constraint that only a single user can be assigned to any tone. The last constraint makes the resource allocation problem an Integer Program (IP) that has been shown to be NP-hard [2]. By making two approximations, the proposed solution reduces this Integer Program (IP) into a Geometric

²In practical systems, when using practical constellations like QAM, the rate expression in Eqn (4) is modified as $s_{km}^i = \log\left(1 + \frac{\gamma_{km}^i P_{km}^i}{G_{QAM}}\right)$, where G_{QAM} denotes the gap to capacity.

³A peak power constraint for each tone arising from a PSD mask can also be easily incorporated.

Program (GP) that can be easily solved with a standard convex optimization software package.

Consider Queue Proportional Scheduling (QPS) for the BE user objective, the IP optimization problem is:

$$\text{maximize } \theta \quad (5)$$

subject to:

$$b_{km} \in \{0, 1\}, \quad \forall k, \forall m \quad (6)$$

$$\sum_k b_{km} = 1, \quad \forall m \quad (7)$$

$$\sum_m \sum_{i=1}^{|\Gamma_{km}|} s_{km}^i P_{km}^i = \theta Q_k^{\text{be}}, \quad \forall k \in \mathcal{K}^{\text{be}} \quad (8)$$

$$\sum_m \sum_{i=1}^{|\Gamma_{km}|} s_{km}^i P_{km}^i = R_k^{\text{req}}, \quad \forall k \in \mathcal{K}^{\text{ds}} \quad (9)$$

$$s_{km}^i \geq 0, \quad \forall k, \forall m, \forall i \quad (10)$$

$$\sum_k \sum_m b_{km} \sum_{i=1}^{|\Gamma_{km}|} \frac{e^{s_{km}^i} - 1}{\gamma_{km}^i} P_{km}^i \leq \bar{P} \quad (11)$$

The variables in the above optimization problem are θ , b_{km} and s_{km}^i . The objective maximizes the BE users' expected-rate via the scalar θ as seen in (5), while constraint (8) requires BE users' rates to be proportional to their queue states. The constraints (6) and (7) on b_{km} limit a tone to be assigned to one and only one user. The constraint (9) forces the DS users' assigned-rates to meet their required rates. The constraint (10) is the non-negativity of rates s_{km}^i . The last constraint (11) requires the power on the assigned tones to be no more than the average available power \bar{P} .

To construct a GP from the IP described by (5)-(11), two approximations and a change of variables are made:

(1) Relax (6) to $0 \leq \rho_{km} \leq 1$. This results in a non-combinatorial problem, however it is not jointly convex.

(2) Define $\tilde{s}_{km}^i = \log(1 + \gamma_{km}^i \rho_{km} P_{km}^i)$. Replace s_{km}^i with \tilde{s}_{km}^i . This results in a jointly convex problem, however it is not in a standard form (e.g. LP, GP).

(3) Approximate \tilde{s}_{km}^i by $r_{km}^i = \log(\gamma_{km}^i \rho_{km} P_{km}^i)$. This results in a GP formulation.

The relaxation (1) and change of variable (2) methods are fairly standard and were also used in [2]. The approximation (3) is fairly accurate in the high signal-to-noise ratio (SNR) regime.

The resulting Geometric Program for QPS⁴ is:

$$\text{maximize } \theta \quad (12)$$

subject to:

$$0 \leq \rho_{km} \leq 1, \quad \forall k, \forall m \quad (13)$$

$$\sum_k \rho_{km} \leq 1, \quad \forall m \quad (14)$$

$$\sum_m \sum_{i=1}^{|\Gamma_{km}|} r_{km}^i P_{km}^i = \theta Q_k^{\text{be}}, \quad \forall k \in \mathcal{K}^{\text{be}} \quad (15)$$

$$\sum_m \sum_{i=1}^{|\Gamma_{km}|} r_{km}^i P_{km}^i = R_k^{\text{req}}, \quad \forall k \in \mathcal{K}^{\text{ds}} \quad (16)$$

$$r_{km}^i \geq 0, \quad \forall k, \forall m, \forall i \quad (17)$$

$$\sum_k \sum_m \frac{1}{\rho_{km}} \sum_{i=1}^{|\Gamma_{km}|} \frac{e^{r_{km}^i}}{\gamma_{km}^i} P_{km}^i \leq \bar{P} \quad (18)$$

The resulting variables in the optimization problem are θ , ρ_{km} and r_{km}^i . The optimal ρ_{km}^* are fractional and must be resolved to binary values before the final allocation is complete. As DS users have a rate requirement R_k^{req} , they have preference over the BE users in selecting tones. While the BE users do not have a hard rate-requirement, they are expecting R_k^{alloc} from the solution of the GP:

$$R_k^{\text{alloc}} = \left(\frac{\sum_{m \in \mathcal{I}_{\text{be}}} \rho_{km}^*}{M} \right) \theta^* Q_k^{\text{be}}, \quad \forall k \in \mathcal{K}^{\text{be}} \quad (19)$$

The first term in Eqn (19) is a scaling factor of the allocated rates based on the fraction of tones assigned to the BE users. The DS users are indexed from 1 through K^{ds} and the BE users are indexed from $K^{\text{ds}} + 1$ to $K^{\text{ds}} + K^{\text{be}}$. Considering R_k^{req} and R_k^{alloc} , the following algorithm assigns tones to the DS and BE users:

Tone Assignment Algorithm :

- (a) $\mathcal{T} := \mathcal{M}, \mathcal{T}_{\text{ds}} := \phi$
- (b) for $k=1$ to K^{ds} ;
- (c) while $R_k^{\text{req}} > 0$;
- (d) if $\mathcal{T} = \phi$, stop;
- (e) $c := \text{argmax}_{m \in \mathcal{T}} \rho_{km}^*$;
- (f) $a_{kc}^* := 1; a_{[k', c, k' \neq k]}^* := 0; \mathcal{T} := \mathcal{T} \setminus c; \mathcal{T}_{\text{ds}} := \mathcal{T}_{\text{ds}} \cup c$;
- (g) $R_k^{\text{req}} := R_k^{\text{req}} - \left[\sum_{i=1}^{|\Gamma_{kc}|} \log(1 + \gamma_{kc}^i P_{kc}^{i*}) \right]$;
- (h) end-while;
- (i) end-for;
- (j) $\mathcal{T}_{\text{be}} := \mathcal{M} \setminus \mathcal{T}_{\text{ds}}$
- (k) Re-index the BE users in decreasing order of their Q_k^{be} i.e. $Q_{1+K^{\text{ds}}}^{\text{be}} \geq Q_{2+K^{\text{ds}}}^{\text{be}} \dots \geq Q_{[K=K^{\text{be}}+K^{\text{ds}}]}^{\text{be}}$;
- (l) for $k=1+K^{\text{ds}}$ to $K^{\text{be}} + K^{\text{ds}}$;
- (m) while $R_k^{\text{alloc}} > 0$;
- (n) if $\mathcal{T} = \phi$, stop;
- (o) $c := \text{argmax}_{m \in \mathcal{T}} \rho_{km}^*$;
- (p) $a_{kc}^* := 1; a_{[k', c, k' \neq k]}^* := 0; \mathcal{T} := \mathcal{T} \setminus c$;
- (q) $R_k^{\text{alloc}} := R_k^{\text{alloc}} - \left[\sum_{i=1}^{|\Gamma_{kc}|} \log(1 + \gamma_{kc}^i P_{kc}^{i*}) \right]$;
- (r) end-while;
- (s) end-for;

If the algorithm exits from line (d), some DS users will suffer a non-zero packet-drop rate. After the tones are allocated,

⁴MWMS,LQHP based on [8] can be formulated as a GP in a similar manner and is skipped because of space constraint.

i.e. the binary-valued assignment variables $a_{km}^* \in \{0, 1\}$ are determined, the last step is to assign power to each tone. For the DS users, this is simply P_{km}^{i*} obtained from $r_{km}^{i*}, k \in \mathcal{K}^{\text{ds}}, m \in \mathcal{T}^{\text{ds}}$. For the BE users, there is the possibility that more power is available. Therefore, the assigned power from the GP solution, i.e. $P_{km}^{i*}, k \in \mathcal{K}^{\text{be}}, m \in \mathcal{T}^{\text{be}}$, is scaled such that all available power, P , is used. There are no iterations of the proposed algorithm⁵.

B. Practical Implementation

For a large number of tones and users, solving a GP is impractical given the current state of GP solvers. Therefore a coalition-based solution based on [11] is used for simulations. Instead of solving a single GP with M tones and K users, smaller non-intersecting subsets of tones and users are formed called coalitions. Each coalition is given a fraction of the total power in proportion with their size and rate-requirements. The GP is solved for each coalition. Using the terminology from [11], solving the GP is the equivalent of bargaining for resources. After the first iteration of bargaining occurs, new coalitions are formed based on the Hungarian method, see [11]. The GP is solved over the new coalitions, and this process iterates until convergence. This approach reduces the complexity of solving a large GP but introduces an iterative approach to the problem.

IV. NUMERICAL RESULTS

This sections presents numerical results comparing the performance of QPS and BCHPR for the BE users, in the presence of DS users. $K^{\text{ds}} = 4$ and $K^{\text{be}} = 6$. The DS users' rate requirement, R_k^{req} , and the BE users' queue size (bits), Q_k , are uniformly-distributed random integers between 1 and 10. The channel fading distribution on each tone is i.i.d. Rayleigh with mean $\bar{\gamma}_{km}$. The mean $\bar{\gamma}_{km}, \forall k, \forall m$ is uniformly-distributed as $U[0, \mu]$. Also, $|\Gamma_{km}|, \forall k, \forall m$ is fixed to 2. We look at quantized channel state feedback because practical receivers feedback only a finite number of bits, in this case 1 bit per tone. Average power constraint, \bar{P} , is set to 50 dB. We compare the performance of BE objectives - QPS and BCHPR in the presence of DS users's rate constraints, R_k^{req} . The rate constraints of the DS users are always met by the solution of the proposed GP, so we compare the performance of the BE users only. Each point is an average over 10 runs, where in each run, all the random variables are generated in an i.i.d. manner. As described in Section III-B, a two-user coalition-based algorithm is used instead of solving one GP over M and K , see [11].

Figure 2 studies the effect of increasing μ for every user on each tone. Total number of tones M is 64. Of the 6 BE users, we focus on two of them - the user with the best channel gain and the user with the longest queue size. As seen from the plot, BCHPR strictly favors the user with the best channel gain (dashed-curve with \square) which clearly maximizes the BE users' sum-rate. However, this strategy is not fair, e.g. a BE

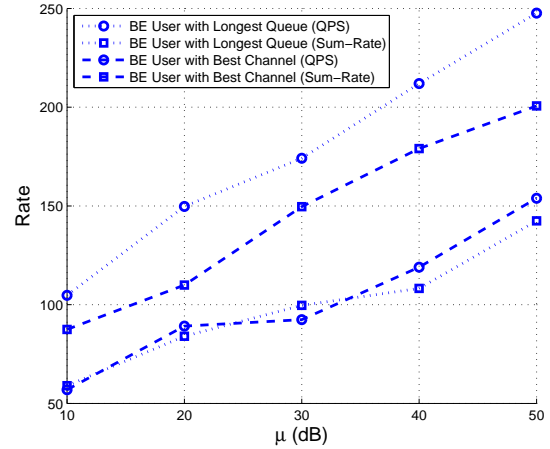


Fig. 2. QPS v/s BCHPR (Sum-Rate Maximization) for the BE users w.r.t. μ .

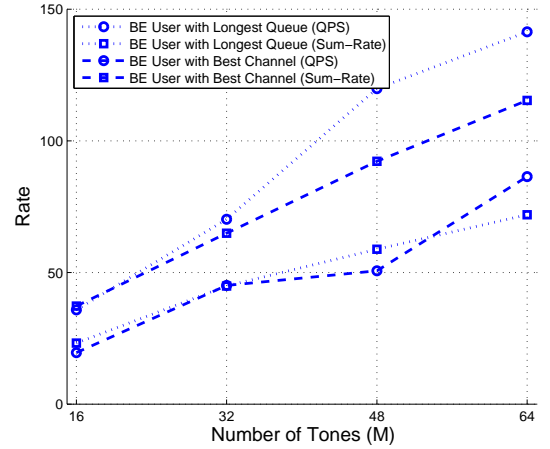


Fig. 3. QPS v/s BCHPR (Sum-Rate Maximization) for the BE users w.r.t. M .

user with a high-queue length and low channel-gain is assigned a very low-rate by BCHPR (dotted-curve with \square). While QPS results in a total rate less than BCHPR, it consider both queue-size (explicitly by maximizing θ Eqn (12)) and channel-gains (implicitly by trying to minimize power-consumption and meeting the \bar{P} constraint), and therefore is considerably more fair than BCHPR. As can be seen from Figure 2, QPS allocates a high-rate to the user with the largest queue-size (dotted-curve with \circ) at the same time assigning a decent rate to the user with the highest channel gain (dashed-curve with \circ). Figure 3 makes the same comparison by increasing M , the total number of tones when μ is fixed to be 20 dB.

Figure 4 shows that the number of iterations for the resource allocation to converge for both the QPS and BCHPR objectives is between 2 and 6.

⁵A similar assignment method is proposed by [3] to force ρ_{km}^* 's to be binary-valued.

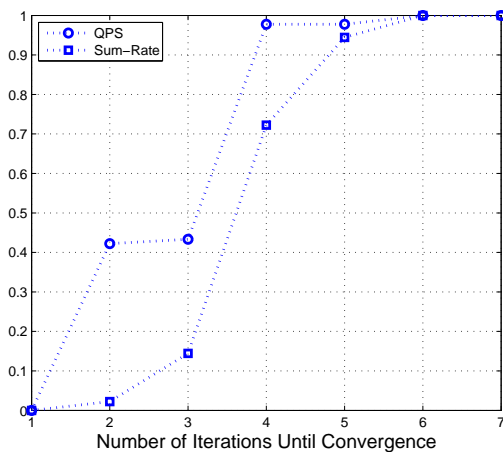


Fig. 4. CDF of Number of Iterations required.

V. CONCLUSION

This paper proposed a persistent-scheduling algorithm for the OFDMA downlink that provides a QoS-guarantee for deadline-sensitive service and efficient resource-allocation for the best effort service. The optimal resource - tone, transmit power and rate allocation problem was formulated as a Geometric Program that can be solved by standard optimization software. An iterative-bargaining approach to solve smaller GPs was provided and used to generate the numerical results, based on the work of [11]. A comparison of two different BE user objectives, QPS and BCHPR, was provided to show the flexibility of the proposed algorithm. By allowing a general-objective solution, a network-designer has the ability to specify an allocation strategy based on a system-level design criteria.

REFERENCES

- [1] Ahmed K. F. Khattab and Khaled M. F. Elsayed, "Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks," *IEEE Computer Society*, 2006.
- [2] C. Wong, R. Cheng, K. Letaief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1747–1758, Oct. 1999.
- [3] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, pp. 362–370, Dec. 2003.
- [4] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for ofdma downlink systems," in *IEEE ISIT*, July 2006.
- [5] J. Huang, V. Subramanian, R. Agarwal, and R. Berry, "Downlink scheduling and resource allocation for ofdm systems," in *CISS*, Mar 2006.
- [6] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," *Proc. IEEE Globecom*, 2000.
- [7] I. C. Wong and B. L. Evans, "Optimal ofdma resource allocation with linear complexity to maximize ergodic weighted sum capacity," in *IEEE ICASSP*, Apr 2007.
- [8] K. Seong, R. Narasimhan, and J. Cioffi, "Queue proportional scheduling via geometric programming in fading broadcast channels," *IEEE Jour. Sel. Areas in Comm.*, vol. 24, pp. 1593–1602, Aug 2006.
- [9] "IEEE 802.16 WiMax Standard."
- [10] R. Agarwal, V. Majjigi, R. Vannithamby, and J. M. Cioffi, "Efficient scheduling for heterogeneous traffic in downlink," in *IEEE VTC Fall*, 2007.

- [11] Z. Han, Z. Ji, and K. R. Liu, "Fair multiuser channel allocation for OFDMA networks using nash bargaining solutions and coalitions," *IEEE Trans. on Communications*, vol. 53, pp. 1366–1376, Aug 2005.