

Low Complexity Resource Allocation with Opportunistic Feedback over Downlink OFDMA Networks

Rajiv Agarwal, *Student Member, IEEE*, Vinay R. Majjigi, Zhu Han, *Member, IEEE*, Rath Vannithamby, and John M. Cioffi, *Fellow, IEEE*

Abstract—Optimal tone allocation in downlink OFDMA networks is a non-convex NP-hard problem that requires extensive feedback for channel information. In this paper, two constant-complexity limited-feedback algorithms are proposed to achieve near-optimal performance. First, using opportunistic feedback, the proposed schemes are shown to reduce feedback overhead by requiring only users likely to be allocated resources to feed back. There are differences between the two proposed schemes for implementation of the feedback protocol. One scheme requires less feedback but is contention-based, while the other scheme is sequential and thus avoids possible collisions leading to slightly higher performance, but needs more feedback. Second, complexity is reduced for resource allocation by solving the optimization problem in a distributed manner, rather than centrally at the base station. As shown both analytically and through numerical results, these distributed algorithms reduce the required feedback overhead significantly, and achieve constant computational complexity with little performance loss compared to the optimal solution.

Index Terms—OFDMA downlink system, Resource allocation, Limited CSIT, Opportunistic feedback, Low transmit complexity.

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 Orthogonal Frequency Division Multiple Access (OFDMA) is an attractive downlink transmission technique. To utilize limited system resources in an efficient manner, *channel-* and *QoS-aware* resource allocation for the OFDMA downlink has played a key role and has attracted much research attention [1]-[4], and references therein.

One of the major problems in employing a resource allocation scheme in OFDMA networks is the computational complexity of the resource allocation problem at the Base Station

Manuscript received December 05, 2007; revised June 20, 2008. Part of this work was presented at the IEEE Global Communications Conference in Nov. 2007.

Rajiv Agarwal (corresponding author), Vinay R. Majjigi and John M. Cioffi are with the Department of Electrical Engineering, Stanford, Stanford, CA 94305, USA (e-mail: {rajivag, vmajjigi, cioffi}@stanford.edu).

Zhu Han is with the Department of Electrical and Computer Engineering, Boise State University, ID 83725, USA (e-mail: zhuhan@boisestate.edu).

Rath Vannithamby is a research scientist at Intel Corp., Hillsboro, OR 97124, USA (e-mail: rath.vannithamby@intel.com).

Digital Object Identifier 10.1109/JSAC.2008.081012.

(BS) because optimal resource allocation in the OFDMA downlink is an NP-hard problem [2] due to the combinatorial search needed for optimal tone allocation. Another problem is the large amount of feedback required to pass to the BS for perfect Channel State Information at the Transmitter (CSIT). Moreover, the frequency of information collocation has to be quicker than that of channel variations. Most works in the literature have addressed these issues individually, however, any OFDMA resource allocation algorithm must consider both computational complexity and feedback overhead when evaluating system performance. Particularly, [5] proposes a low-complexity algorithm for solving the resource allocation problem in the dual domain when perfect CSIT is available. This scheme is attractive under settings when the channel variations are infrequent (e.g. DSL) or when perfect CSIT can be easily made available (e.g. Time Division Duplexing (TDD) setup). However, in an Frequency Division Duplexing (FDD) setup with relatively short channel coherence time (e.g. 802.16 WiMAX), the feedback overhead for perfect CSIT can be overwhelming. For example in 802.16 WiMAX, there are a total of 2048 tones, so for a network with 100 users, the BS needs to collect 204,800 real numbers for perfect CSIT.

Reducing feedback overhead has thus attracted much research attention [6]-[17]. The authors in [7]-[11] discuss reducing feedback overhead for a Multiple Input Multiple Output (MIMO) system, [12]-[13] describe transmission techniques to tackle limited feedback in a single-user Orthogonal Frequency Division Multiplexing (OFDM) system and [14]-[17] deal with a multi-user OFDMA system. Schemes proposed in [14]-[17] reduce the feedback overhead per user, however, all users feed back to the BS.

A way to reduce system-wide feedback overhead is by exploring opportunistic feedback only from the users that are most likely to be allocated resources [18]-[23]. The basic idea in all Opportunistic Feedback (OF) protocols is to let users access the feedback medium opportunistically based on the value of their channel-gains. There are pre-defined thresholds for channel-gain, and the feedback slots are associated with these thresholds. In any feedback slot, users compare their channel-gain values with the pre-defined threshold and transmit to the BS if their channel-gain value is above the threshold. If the objective of resource allocation is to maximize sum-rate, the problem of resource allocation in the OFDMA downlink reduces to identifying the user with the highest

channel value for each tone. An opportunistic splitting scheme is proposed in [18], where a pair of thresholds is initially chosen by the BS and broadcast to all the users. This pair of thresholds is iteratively updated and broadcast again until only (the user with) the highest channel gain lies between the two thresholds. OF protocols using a *single* threshold are proposed in [19]-[21], where the value of the threshold is optimized. In [19], users are divided into multiple groups, each group corresponding to a separate feedback slot. In each slot, users assigned to that slot contend if their channel gain value is above the threshold. In [20], all users can contend in any feedback slot, however, they do so with some probability to reduce contention. This probability is optimized along with the threshold. The authors in [21] study the case of finite coherence time and optimize the threshold along with a few other system parameters to maximize sum-rate normalized by the fraction of time available for data-transmission.

This paper extends our work in [22]-[23] and uses *multiple* thresholds. Further, this paper addresses both feedback overhead and transmitter complexity. Two constant-complexity, limited-feedback algorithms are proposed to achieve near-optimal performance. Both proposed algorithms use the Lagrange decomposition method to solve the Weighted Sum Rate Maximization (WSRmax) problem with two contributions. First, the proposed feedback schemes are based on OF and so reduce the feedback overhead. The difference between the two proposed schemes arises from the way the feedback protocol is implemented. One scheme requires less feedback but is contention-based, while the other scheme is sequential and thus avoids possible collisions leading to slightly higher performance, however with comparatively higher feedback. The second contribution of this paper is to reduce transmitter complexity in performing resource allocation, which is achieved by solving the optimization problem in a distributed manner, rather than centrally at the BS. The Lagrange dual method requires a bisection search of the dual variable to solve the optimization problem. In the proposed two schemes, this bisection search is performed through multiple interactions between the users and the BS rather than solely at the BS. The WSRmax problem is studied because in practice different users may have different priorities arising from different QoS requirements. As shown both analytically and through numerical results, these distributed algorithms reduce the required feedback significantly and achieve *constant* computational complexity with little performance loss compared to the optimal solution. The performance of the proposed schemes is compared for sum-rate with schemes proposed in [19], [20], which were designed for maximizing sum-rate, numerically. The proposed schemes are shown to give higher sum-rate with a lower feedback overhead as compared to the scheme in [19]. The scheme in [20] is found to give very low sum-rate, though with very low feedback overhead.

The paper is organized as follows: Section II describes the system model and the WSRmax problem formulation, followed by a discussion on solving the WSRmax problem in two ways. The first is the optimal solution and the second is the solution proposed in [5]. Two OF protocols are then described in Section III as proposed solutions to the WSRmax problem that utilize the Lagrange dual decomposition idea

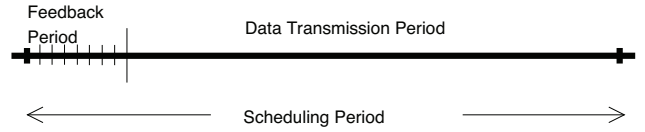


Fig. 1. Scheduling Period Partition.

from [5]. Finally, Section IV presents the numerical results, and Section V provides concluding remarks.

II. SYSTEM MODEL, PROBLEM FORMULATION, AND EXISTING SCHEME

Consider a downlink transmission system with K users and M tones where the BS and each user are equipped with a single omnidirectional antenna. For any given power delay profile at the users, it is assumed that the use of OFDM technique eliminates any Inter-symbol Interference (ISI), resulting in M non-interfering independently-fading parallel channels in the frequency domain. The total downlink transmit power is constrained to P_{tot} . For user k on tone m , the channel value is denoted by h_{km} , which is assumed to be perfectly known at the respective receivers. It is assumed the channel values for all users are independent and identically distributed (i.i.d.) for all tones¹. A zero-mean i.i.d. Gaussian noise with variance σ_{km}^2 is added at the receiver part. The channel Signal-to-Noise Ratio (SNR) for user k on tone m is defined as $\gamma_{km} \triangleq \frac{|h_{km}|^2}{\sigma_{km}^2}$. The long-term statistics i.e. the probability density function (pdf) and cumulative distribution function (cdf) of γ_{km} are denoted by $f_{\Gamma}(\gamma)$ and $F_{\Gamma}(\gamma)$ respectively and are assumed to be known to the users as well as the BS. We define a Scheduling Period (SP) to consist of a feedback period plus a data transmission period. This is depicted in Figure 1. We assume that for each tone m , γ_{km} , the channel gain of user k , follows a block-fading model where γ_{km} remains constant for an SP and then changes in an i.i.d. manner to a new realization.

The concern of this paper is dynamic allocation of power and rate on each tone based on the knowledge of channel conditions at the BS. Let \mathcal{S}_k denote the set of tones allocated to user k . Each tone is allowed to be used by at most one user; hence, $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for $i \neq j$ and $\cup_{k=1}^K \mathcal{S}_k \subseteq \{1, 2, \dots, M\}$. The transmitter finds \mathcal{S}_k for all $k = 1, 2, \dots, K$ and distributes power so that the objective of resource allocation is satisfied.

Let r_{km} and p_{km} denote the rate and the power of user k on tone m such that $r_{km} = \log(1 + p_{km}\gamma_{km})$. If we define w_k as the weight assigned to user k , given the weight vector and the channel gains, the WSRmax problem finds the tone and power allocation that maximizes the weighted sum rate with the total power constraint, i.e.

$$\max_{p_{km}, \mathcal{S}_k} \sum_{k=1}^K w_k \sum_{m \in \mathcal{S}_k} r_{km} \quad (1)$$

¹The i.i.d. assumption is needed for keeping the analysis tractable. The proposed OF protocols, however, can be easily extended for the non-i.i.d. case following the discussion in Section III-C.

$$\text{subject to } \begin{cases} \sum_{k=1}^K \sum_{m \in \mathcal{S}_k} p_{km} \leq P_{\text{tot}}, \\ \mathcal{S}_i \cap \mathcal{S}_j = \phi, \quad \forall i \neq j, \\ \cup_{k=1}^K \mathcal{S}_k \subseteq \{1, 2, \dots, M\}, \\ p_{km} \geq 0, \quad \forall k \text{ and } \forall m. \end{cases}$$

The boundary of the achievable rate region can be traced by solving this problem for all possible weight vectors. In general, (1) is not a convex optimization problem. Moreover, (1) is NP-hard since it needs to find the optimal set of tones for each user, which is a combinatorial problem whose complexity increases exponentially with M . Next, we discuss two schemes for resource allocation in the OFDMA downlink. Their description will be useful for the development of the proposed scheme in Section III. Both schemes require perfect Channel State Information (CSI) at the transmitter.

A. Scheme-I: Full CSI and Solving the WSRmax Problem

In order to solve (1), Scheme-I tries all possible combinations of allocations of tones to users and chooses the allocation that maximizes the weighted-sum-rate. The complexity grows exponentially with the number of tones M . This is because each tone can be allocated to any of the K users and hence $K \times K \times \dots \times K = K^M$ possible allocations. Once an allocation has been made, power assignment on each tone has a closed-form expression arising from water-filling [26], which has constant computational complexity, independent of both K and M . Hence, the complexity of Scheme-I is $\mathcal{C}_I = O(K^M)$. The performance or the objective function value \mathcal{V}_I of Scheme-I is the highest as it solves (1) optimally. Thus \mathcal{V}_I serves as an upper-bound for the performance of any other scheme.

Since this scheme requires perfect CSI, the feedback overhead incurred by Scheme-I is $\mathcal{F}_I = MKB_{\text{real}}$, where B_{real} is the number of bits used for quantizing the real numbers corresponding to the channel-gain γ_{km} values. B_{real} is fixed and is a system design parameter.

B. Scheme-II: Full CSI and Solving the WSRmax Problem in the Dual-Domain

Next, we formulate the dual function of (1) in a similar manner as done in [5]. The first step is to form the Lagrangian or the dual function of WSRmax problem in (1) over a domain \mathcal{D} as

$$\mathcal{L}(\{p_{km}\}, \lambda) = \sum_{k=1}^K w_k \sum_{m=1}^M r_{km} - \lambda \left(\sum_{k=1}^K \sum_{m=1}^M p_{km} - P_{\text{tot}} \right), \quad (2)$$

where the domain \mathcal{D} is defined as the set of all non-negative p_{km} 's for $k = 1, 2, \dots, K$ and $m = 1, 2, \dots, M$ such that for each tone m , at most one p_{km} is positive for $k = 1, 2, \dots, K$. The summations in Eqn (2) can be rearranged as

$$\mathcal{L}(\{p_{km}\}, \lambda) = \sum_{m=1}^M \left[\sum_{k=1}^K w_k r_{km} - \lambda \left(\sum_{k=1}^K p_{km} \right) \right] + \lambda P_{\text{tot}}. \quad (3)$$

Then, the Lagrange dual function is

$$g(\lambda) = \max_{\{p_{km}\} \in \mathcal{D}} \mathcal{L}(\{p_{km}\}, \lambda). \quad (4)$$

It was shown in [5] that the above problem can be decomposed into M sub-problems, one for each tone and can be solved for each tone m separately.

In summary, Scheme-II fixes a value of λ and for each tone m , finds the user that contributes the maximum to the WSRmax objective by maximizing the objective for each tone as

$$h_m(\lambda) = \max_k \left\{ w_k \log \left(1 + \left(\frac{w_k}{\lambda} - \frac{1}{\gamma_{km}} \right)_+ \gamma_{km} \right) \right\}, \quad \forall m, \quad (5)$$

and then iterates over the value of λ until the power constraint is satisfied. The user k_m^* who is assigned to tone m is the one that maximizes the contribution to the WSRmax objective from tone m as given in (5). Thus

$$k_m^* = \arg \max_k \left\{ w_k \log \left(1 + \left(\frac{w_k}{\lambda} - \frac{1}{\gamma_{km}} \right)_+ \gamma_{km} \right) \right\}. \quad (6)$$

A special case is of particular interest, when $w_k = \frac{1}{K}, \forall k$ i.e., the objective is to maximize throughput or sum-rate. In this special case, the equations are simplified significantly. The computation of the best user for every tone is simplified as

$$k_{m,\text{sr}}^* = \arg \max_k \left\{ \log \left(1 + \left(\frac{1}{\lambda'} - \frac{1}{\gamma_{km}} \right)_+ \gamma_{km} \right) \right\} \quad (7)$$

$$\stackrel{(a)}{=} \arg \max_k \left\{ \log \left(1 + \left(\frac{1}{\lambda'} - \frac{1}{\gamma_{km}} \right) \gamma_{km} \right) \right\} \quad (8)$$

$$\stackrel{(b)}{=} \arg \max_k \left\{ \log \left(\frac{\gamma_{km}}{\lambda'} \right) \right\} \quad (9)$$

$$\stackrel{(c)}{=} \arg \max_k \gamma_{km}. \quad (10)$$

We defined λ' to distinguish from the λ used for the WSRmax problem in (1). The expressions in (a) and (b) follow from the understanding that the maximization is performed only over those users for whom the power $\left(\frac{1}{\lambda'} - \frac{1}{\gamma_{km}} \right)$ is non-negative and if none of them satisfy it, the tone is left unallocated. (c) follows from observing that $\log \left(\frac{\gamma_{km}}{\lambda'} \right)$ is a monotonically increasing function of γ_{km} for a fixed positive λ' . This is intuitive, as for maximizing throughput, any tone (if used) is given to the user with the highest channel gain.

The number of iterations required for convergence of λ (or λ'), denoted by N_{iter} , depends on the required accuracy and is independent of both K and M [27]. Hence, the complexity of Scheme-II comes from solving (5) only, which has a complexity $\mathcal{C}_{\text{II}} = O(N_{\text{iter}}MK)$. Notice that as compared to \mathcal{C}_I of Scheme-I, the complexity of Scheme-II is linear in both M and K . Since Scheme-II requires perfect CSI as well, the feedback overhead $\mathcal{F}_{\text{II}} = \mathcal{F}_I = MKB_{\text{real}}$ is the same as that of Scheme-I. Since the original problem (1) is not convex, there is a duality gap between the solution of the original and the dual problem being solved by Scheme-II. The duality gap is non-zero and so the value of the objective function of Scheme-II, \mathcal{V}_{II} , is less than the maximum possible. However, the duality gap has been shown to be negligible for moderate values of M ($M \geq 8$) and have been proved to go to zero as M goes to infinity [25]. For our simulations in Section IV, we use values $M = 52$ and 2048, hence for these values, $\mathcal{V}_I = \mathcal{V}_{\text{II}}$.

III. PROPOSED PROTOCOL: PARTIAL CSI AND SOLVING THE DUAL PROBLEM

The proposed scheme aims to reduce both feedback-overhead and transmitter-complexity by modifying Scheme-II. The main idea for the proposed scheme comes from the definition of $h_m(\lambda)$ in (5). As seen in (5), any tone is allocated to a single user, that maximizes the per-tone objective function. Thus, the only channel gain-value required at the BS for power computation is the channel-gain value of the user who gets the tone. The BS needs to find the best user w.r.t. the per-tone objective function (5) and then the actual channel-gain value of the best user only. The problem, thus reduces to finding the maximum of a collection of random variables in a distributed manner. This is what the OF protocol achieves.

For simplicity, we first consider the sum-rate case i.e. when $w_k = \frac{1}{K}$. In this case, (6) simplifies to (10) and the BS just needs to know the best user k_m^* and its channel-gain $\gamma_{k_m^* m}$ for every tone m . The OF protocol, which lets users with higher channel-gain values to feedback first, thus can be used to solve (10). In this paper, two versions of the OF protocol are proposed: one using contention-based feedback and the other using sequential-feedback. The feedback overhead $\mathcal{F}_{\text{III}}^{\text{con}} = (c_1 M \log K)$ and $\mathcal{F}_{\text{III}}^{\text{seq}} = (c_2 M \log M \log K)$ of the two schemes are evaluated analytically, where c_1 and c_2 are constants. Notice that if the BS knows k_m^* and $\gamma_{k_m^* m}$ for every tone m , the only part of resource allocation left is to water-fill power, which has a closed-form expression. Hence the complexity of resource-allocation $\mathcal{C}_{\text{III}}^{\text{con}} = \mathcal{C}_{\text{III}}^{\text{seq}}$ for both the proposed schemes is $O(1)$, which is independent of both M and K . The performance $\mathcal{V}_{\text{III}}^{\text{con}}$ and $\mathcal{V}_{\text{III}}^{\text{seq}}$ are shown to be ‘close’ to \mathcal{V}_{II} . Here, the term ‘close’ is made precise by deriving an upper-bound on $(\mathcal{V}_{\text{II}} - \mathcal{V}_{\text{III}}^{\text{con}})$ and $(\mathcal{V}_{\text{II}} - \mathcal{V}_{\text{III}}^{\text{seq}})$.

In Section III-A, we describe the first proposed scheme, a contention-based OF protocol, and analyze its performance. This scheme is an extension of the scheme proposed by us in an earlier work [23]. The description of the scheme is given in detail for two reasons - (a) sake of completeness and (b) to serve as a building-block for the description of the second proposed scheme. However, some of the proofs in the analysis carry-over and so are duly cited. In Section III-B, we describe and analyze the second OF protocol based on sequential feedback. Both the schemes are described for solving (10) or the sum-rate maximization problem. In Section III-C, we describe the modification for the two proposed protocols to solve (6) and consequently the original WSRmax problem in (1).

A. Contention-Based OF Protocol

1) *Operation of the Protocol:* The operation of the contention-based OF protocol for a single tone (say tone m) is shown in Figure 2. For M tones, the same protocol is repeated M times in sequence. Before the feedback-period begins, the BS broadcasts a training sequence that allows all users to estimate their channel gains γ_{km} for tone m . Subsequently, the BS broadcasts an *Initiate Feedback* message, which starts the feedback-period. During the feedback-period for tone m , there are L feedback slots and these slots are associated with thresholds $T_0 > T_1 > T_2 > \dots > T_L$. In the i^{th} ($i = 1, 2, \dots, L$) feedback slot, any user k transmits a

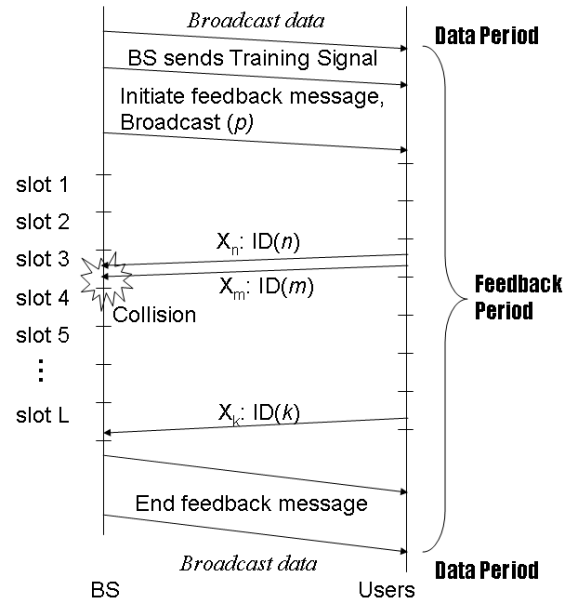


Fig. 2. The Proposed Contention-Based OF Protocol for a Single Tone.

message x_k^{con} to the BS if its channel gain γ_{km} satisfies the relation $T_i \leq \gamma_{km} < T_{i-1}$. Since the threshold values are in a decreasing order, the first feedback-slot in a feedback-period is accessed by the users with the highest channel-gain on the tone. The message x_k^{con} for any user contains the unique user number k , for which the number of bits required is $\log K$. If the feedback message x_k^{con} , sent by the user who is the first one to transmit in the feedback period, is received successfully, the BS knows k_m^* and $\gamma_{k_m^* m}$. The BS knows k_m^* from the unique user identity k contained in the message x_k^{con} and it has an estimate of $\gamma_{k_m^* m}$ given by the lower threshold of the feedback slot T_j , where j is the feedback slot in which the feedback message by this user x_k^{con} is transmitted. The threshold values are the same for all k and m , and are computed once and fixed for all time.

Depending upon the channel-gain value in the current SP, many users may transmit their x_k^{con} in the same feedback slot. Due to the contention between the users, this scheme is called the contention-based OF protocol. If two or more users happen to feedback their x_k^{con} 's in the same feedback slot, their feedback messages will collide and are assumed irrecoverable². In case of a collision in a feedback slot, the feedback-period continues on subsequent feedback slots and ends after L feedback slots, when the BS broadcasts an *End Feedback* message. If no user's x_k^{con} is received successfully in the L slots, the tone is left unallocated, else it is allocated to the first user that transmitted successfully.

2) *Calculation of Thresholds:* The set of thresholds T_i are determined by partitioning the pdf of the channel gain $f_{\Gamma}(\gamma)$

²In [20], it is also assumed that if in a feedback slot, more than one user transmit together leading to a collision, the messages (containing the user identities) cannot be decoded and the BS selects a random user from the entire user pool for scheduling. In another existing work [19], the users that would contend for any feedback slot are known to the BS apriori, hence after a collision, one of the collided users can be selected for scheduling. In this work, the assumption that collided messages cannot be decoded is made to keep the description of the proposed protocol simple. Additionally, it also simplifies the analysis of the protocol.

into N equiprobable intervals, as also done in [11], [23]. The probability that user k transmits its feedback message x_k^{con} for tone m in the i^{th} slot is

$$P\{T_i < \gamma_{km} < T_{i-1}\} = F_{\Gamma}(T_{i-1}) - F_{\Gamma}(T_i), i = 1, 2, \dots, N. \quad (11)$$

To partition the pdf, the probability that a user transmits a feedback message for a tone in any slot is set to the same value $p = \frac{1}{N}$ by choosing thresholds such that $F_{\Gamma}(T_{i-1}) - F_{\Gamma}(T_i) = p$, implying

$$T_i = F_{\Gamma}^{-1}(1 - ip), i = 1, \dots, N; T_0 \triangleq \infty. \quad (12)$$

The parameter p is a design parameter that affects the probability with which any user transmits in a feedback slot and is described in Section III-A3.

3) *Analysis*: To analyze the proposed protocol, the number of users that can transmit in any feedback slot is assumed to be the same for all slots until a successful transmission takes place, and this assumption is ratified with a probabilistic argument. In reality, the number of users that can feedback in any feedback slot may decrease as the feedback protocol proceeds because in an earlier feedback slot, users may have collided; however as we argue next, the probability of collision can be made very small. Since the users have i.i.d. fading, the probability that any user transmits in a given feedback slot i is p . This probability p is the same for all users and for all feedback slots. So, transmission during a feedback slot can be modeled as a binomial random variable with parameters K and p , which is simply the sum of K i.i.d. Bernoulli random variables with parameter p . The probabilities that zero and one user transmit in a feedback slot are q^K and $r = Kpq^{K-1}$ respectively. When more than one user transmit in a slot, they will collide and the probability of this collision is $c = 1 - r - q^K$. Lemma 1 shows that c can be made to go to zero by design.

Lemma 1: If $q = s^{1/K}$, c is upper-bounded by a constant that goes to 0 if the design parameter s is chosen close to 1.

Proof: If $q = s^{1/K}$, then

$$c = 1 - r - q^K \stackrel{(a)}{=} 1 - Kpq^{K-1} - s \quad (13)$$

$$\stackrel{(b)}{=} 1 - K \left(1 - s^{1/K}\right) \left(s^{\frac{K-1}{K}}\right) - s \quad (14)$$

$$\lim_{K \rightarrow \infty} \leq 1 - s \left[1 + \ln \frac{1}{s}\right] \quad (15)$$

$$\stackrel{=}{=} \lim_{s \rightarrow 1} 0, \quad (16)$$

where (a) and (b) follow by substitution. As seen from (14), c is a monotonically increasing function of K , upper-bounded by a constant given in (15). The value of this constant can be made to close to 0 by choosing s close to 1. \square

Under the assumption that the number of users that can transmit in any feedback slot is the same for all slots (valid when $s \rightarrow 1$), the number of feedback slots till successful reception, denoted by a random variable S , is geometrically distributed with probability r , thus $\mathbb{E}[S] = \frac{1-r}{r}$.

Theorem 1: $\mathbb{E}[S]$ is upper-bounded by a small positive number i.e. $\mathbb{E}[S] \leq \frac{1}{s(1-s)}$. Notice that s is independent of K and hence the upper bound $\frac{1}{s(1-s)}$ is also independent of K .

Proof: Given in [23].

The result of Theorem 1 casts light on the choice of the value of design parameter s and its physical significance. If the

value of s is very close to 1, say $s = 0.99$, the probability of transmission p for a $K = 100$ user system is very small, only 10^{-4} . Hence, most feedback slots will be idle; collisions will be rare (probability of collision $c = 5 \times 10^{-5}$ is negligible). So, the user with the highest channel gain will be found with a very high probability $(1-c)$. However, the expected number of feedback slots until we get a successful reception, $\frac{1}{(1-s)s} = 101$, is large. On the other hand, if $s = 0.9$, $p = 10^{-3}$ and $c = 5 \times 10^{-3}$. There are less idle slots and the feedback period finishes sooner $\frac{1}{(1-s)s} = 11$, though the chance of losing the feedback message from the user with the highest channel-gain value due to a collision is higher now. Thus the value of s can be used to trade-off feedback overhead for the rate achieved.

Choosing the number of feedback slots in a feedback period per-tone to be a fixed constant i.e. $L = c_a \frac{1}{(1-s)s}$, $c_a > 1$, ensures that by the end of the feedback period, a feedback message is received successfully with probability greater than $1 - \frac{1}{c_a}$, which follows from a simple use of Markov inequality. By Markov inequality [29], the probability that a non-negative random variable is larger than its expected value by a factor of c_a is less than $\frac{1}{c_a}$. Since $\mathbb{E}[S]$ is at most $\frac{1}{(1-s)s}$, the result follows. The feedback-overhead per tone in bits is $L \log K$ and therefore the total feedback overhead is

$$\mathcal{F}_{\text{III}}^{\text{con}} = c_a \frac{1}{(1-s)s} M \log K, \quad (17)$$

which grows as $M \log K$ only, as opposed to MKB_{real} for both Scheme-I and Scheme-II. Further, the chosen value of L or equivalently c_a is another design parameter that can be used to trade-off feedback overhead for rate achieved on a tone. If L is large, the feedback protocol for any tone can wait longer for the first successful feedback and hence avoid the situation in which the tone is left unallocated. This comes at the cost of longer feedback period or higher feedback overhead.

The expected rate on the m^{th} tone, can be written as

$$\mathbb{E}[r_m^{\text{con}}] = \Pr_m^{\text{alloc}} \sum_{l=0}^{L_{\text{suc}}} (r+s)(1-(r+s))^l \mathbb{E}[\log(1 + \tilde{\gamma}_{[K:K-n_l],m} P_m^{\text{con,wf}})], \quad (18)$$

where $\Pr_m^{\text{alloc}} = \text{Prob}(\text{Tone } m \text{ is allocated})$. The summation is over the unsuccessful transmission attempts, in which two or more users transmit until the $L_{\text{suc}}^{\text{th}}$ successful transmission attempt in which a single user transmitted. $\tilde{\gamma}_{[K:j],m}$ is the random variable characterizing the j^{th} maximum of a set of K i.i.d. random variables $\tilde{\gamma}_{km}$. n_l is the number of users that were lost due to collisions up to the l^{th} unsuccessful transmission attempt, with $n_0 \triangleq 0$. $P_m^{\text{con,wf}}$ is the power assigned to tone m once the feedback period is complete on all M tones, obtained by water-filling P_{tot} over the M tones. Notice that Eqn (18) uses $\tilde{\gamma}$ instead of the actual channel-gain value γ because the channel-gain estimate is obtained from the lower-threshold of the feedback-slot in which the successful transmission takes place. A lower bound on the rate expression in Eqn (18) is

$$\begin{aligned} \mathbb{E}[r_m^{\text{con}}] &\stackrel{(a)}{\geq} \Pr_m^{\text{alloc}} (r+s) \mathbb{E}[\log(1 + \tilde{\gamma}_{[K:K],m} P_m^{\text{con,wf}})] \quad (19) \\ &\stackrel{(b)}{\geq} \Pr_m^{\text{alloc}} s \mathbb{E}[\log(1 + \tilde{\gamma}_{[K:K],m} P_m^{\text{con,wf}})] \quad (20) \\ &\stackrel{(c)}{\geq} \Pr_m^{\text{alloc}} s \mathbb{E}[\log(1 + (\gamma_{[K:K],m} - \Delta_m^{\text{con}}) P_m^{\text{con,wf}})], \quad (21) \end{aligned}$$

where (a) follows by retaining only the first-term in the summation in Eqn (18) (since all terms are non-negative), (b) follows from the fact that both r and s are > 0 , and (c) follows from the definition of Δ_m^{con} provided next. Let Δ_m^{con} be defined as

$$\Delta_m^{\text{con}} \triangleq \max\{\max_{i>0} (T_{i-1} - T_i), \max_k \gamma_{km} - T_1\}. \quad (22)$$

Hence, Δ_m^{con} is the maximum interval between the lower and upper thresholds³ T_{i-1} and T_i , implying that $\gamma_{[K:K],m} - \tilde{\gamma}_{[K:K],m} \leq \Delta_m^{\text{con}}$ always and hence the relation in (c) (21). The value of Δ_m^{con} depends on the choice of parameter s and the channel-gain cdf $F_\Gamma(\gamma)$. Using the lower bound on $\mathbb{E}[r_m^{\text{con}}]$, we come up with the following upper-bound

$$\begin{aligned} \mathcal{V}_{\text{II}} - \mathcal{V}_{\text{III}}^{\text{con}} &\leq \sum_{m=1}^M \mathbb{E} [\log (1 + \gamma_{[K:K],m} P_m^{\text{wf}})] - \left[\Pr_m^{\text{alloc}} s \right. \\ &\quad \left. \mathbb{E} [\log (1 + (\gamma_{[K:K],m} - \Delta_m^{\text{con}}) P_m^{\text{con,wf}})] \right] \\ &\stackrel{(a)}{\leq} \sum_{m=1}^M \mathbb{E} [\log (1 + \gamma_{[K:K],m} P_m^{\text{wf}})] - \left(1 - \frac{1}{c_a} \right) \\ &\quad s \mathbb{E} [\log (1 + (\gamma_{[K:K],m} - \Delta_m^{\text{con}}) P_m^{\text{con,wf}})] \quad (23) \\ &= U_a \quad (24) \end{aligned}$$

where (a) follows from the discussion above i.e. the probability of a tone being allocated $\Pr_m^{\text{alloc}} \geq \left(1 - \frac{1}{c_a}\right)$, P_m^{wf} and $P_m^{\text{con,wf}}$ are the result of the water-filling of power P_{tot} over the channel gains $\gamma_{[K:K],m}$'s and $(\gamma_{[K:K],m} - \Delta_m^{\text{con}})$'s, respectively. As seen through Eqn (23), Δ_m^{con} and consequently the performance gap U_a becomes smaller as c_a increases and/or s gets closer to 1 because both the factors increase L , the feedback period duration⁴.

B. Sequential OF Protocol

1) *Operation of the Protocol:* The operation of the sequential OF protocol is the same as that of the contention-based OF protocol with only two differences: users feedback sequentially and for all M tones at the same time. Before the feedback-period begins, the BS broadcasts a training sequence that allows all users to estimate their channel gains γ_{km} for all tones. Users feedback sequentially during the feedback slot. In the i^{th} feedback slot, first of all, user 1 feeds back an M -length bit vector consisting of zeros and ones corresponding to the M tones. A bit corresponding to a tone is 1 if the channel-gain value on that tone satisfies $T_i \leq \gamma_{km} < T_{i-1}$. Following user 1, the other users (user 2 through user K in sequence) feed back their M -length bit vectors in a similar manner. The bit-vector is compressed and the number of bits after

³The value of Δ_m^{con} additionally depends on the maximum channel realization among all users for tone m in a particular instant. However, in practice (through simulations), we found that the first term $\max_{i>0} (T_{i-1} - T_i)$ dominates the second term $\max_k \gamma_{km} - T_1$ in (22).

⁴It should be noticed that given a finite coherence time, in the absence of channel reciprocity, uplink transmissions during the feedback period come at the cost of a shorter time window available for downlink transmission of actual data (as can be seen from Figure 1) and hence affect data-rate. Hence, although increasing L increases the downlink sum-rate by itself, the effective sum-rate, obtained by scaling the sum-rate expression in Eqn (18) by the fraction of coherence time available for data-transmission, might decrease. In this paper, as also done in [20], [19], we analyze the feedback overhead and the downlink data-rate individually and do not combine the two. Some other works in literature [21], [24] evaluate downlink transmission rate by taking into account time lost due to feedback explicitly.

compression depends on the probability of having 1's at any position in the vector, which is $p = \text{Prob}(T_i \leq \gamma_{km} < T_{i-1})$. Since, users feed back sequentially, there is no collision and the feedback period ends after L feedback slots, at which point the BS broadcasts an *End Feedback* message. The BS allocates a particular tone (say tone m) to the user, who is the first one to have transmitted bit 1 for tone m . If two or more users have transmitted bit-1 for a particular tone in the same feedback slot, the tone is randomly assigned to any of these users. If no user feeds back a 1-bit for a particular tone in all of L feedback slots, the tone is left unallocated.

2) *Analysis:* Since the channel-gain values are i.i.d., p (defined in Section III-A2) denotes the probability that any user feeds back a bit-1 for any tone in any feedback slot. The number of bits that the M -length bit vector can be compressed to is given by $MH(p)$, where $H(p)$ is the standard entropy function defined as $H(p) \triangleq -p \log p - (1-p) \log(1-p)$ [29]. Given this, the number of bits transmitted during any feedback slot in a feedback period is M times

$$KH(p) \stackrel{(a)}{=} K(-p \log p - (1-p) \log(1-p)) \quad (25)$$

$$\stackrel{(b)}{=} \left[K \left(- \left(1 - s^{\frac{1}{K}} \right) \log \left(1 - s^{\frac{1}{K}} \right) - \left(s^{\frac{1}{K}} \right) \log \left(s^{\frac{1}{K}} \right) \right) \right] \quad (26)$$

$$\stackrel{(c)}{=} \log \left(\frac{1}{s} \right) + \log \left(\frac{1}{s} \right) \log K - \log \log \left(\frac{1}{s} \right) \log \left(\frac{1}{s} \right) \quad (27)$$

where (a) follows by the definition of $H(p)$, (b) follows from the definition of $p = 1 - q = 1 - s^{\frac{1}{K}}$ as defined in Section III-A2, and (c) follows from a Taylor-series expansion by replacing $z = \frac{1}{K}$ and then expanding w.r.t. to z around $z = 0$.

The probability that no user transmits for a tone in a feedback slot is q^K . As soon as one or more than one user transmit for a tone (say tone m) in a slot, tone m can be allocated. If $q = s^{1/K}$, the number of feedback slots till we get a bit-1 for tone m , denoted by a random variable S_m , is geometrically distributed with success probability $1 - q^K = 1 - s$, thus $\mathbb{E}[S_m] = \frac{s}{1-s}$. Now, let the number of feedback slots till we get a 1-bit for all M tones be denoted by the random variable S , we have the following theorem.

Theorem 2: For the proposed sequential OF protocol, $\mathbb{E}[S]$ is upper-bounded by $\mathbb{E}[S_m] \log M$.

Proof: From the description of the sequential OF protocol, S is simply $\max_m \{S_1, S_2, \dots, S_M\}$ and the result in the theorem follows from the extreme-value theorem in [28] \square .

By choosing L to be a fixed constant $c_b \frac{s}{(1-s)} \log M$, $c_b > 1$, we can ensure that at the end of the feedback period, all tones have been allocated with probability greater than $1 - \frac{1}{c_b}$, which again follows from a simple use of Markov inequality. Therefore, the total feedback-overhead in terms of bits is

$$\begin{aligned} \mathcal{F}_{\text{III}}^{\text{seq}} &= c_b \frac{s}{(1-s)} M \log M \left(\log \left(\frac{1}{s} \right) + \log \left(\frac{1}{s} \right) \log K \right. \\ &\quad \left. - \log \log \left(\frac{1}{s} \right) \log \left(\frac{1}{s} \right) \right) \quad (28) \end{aligned}$$

and grows as $M \log M \log K$ as opposed to MKB_{real} for both Scheme-I and Scheme-II.

The next step is to find a lower bound of the sum-rate achieved by the sequential OF protocol.

$$\mathbb{E} [r_m^{\text{seq}}] \geq \Pr^{\text{alloc}} \mathbb{E} [\log (1 + (\gamma_{[K:K],m} - \Delta_m^{\text{seq}}) P_m^{\text{seq,wf}})], \quad (29)$$

where $\Pr^{\text{alloc}} = \text{Prob}(\text{All } M \text{ tones are allocated})$ and the inequality follows from the definition of Δ_m^{seq} , which is the same as that given before for Δ_m^{con} (22). Recall, Δ_m^{seq} is the maximum interval between the lower and upper thresholds T_{i-1} and T_i , implying that if for a particular tone (say tone m), more than one user transmit a 1-bit in the same feedback slot, then by assigning the tone randomly to one of these users, the BS can underestimate the channel-gain value for tone m by a maximum of Δ_m^{seq} and hence the relation in (29). Using the lower bound on $\mathbb{E} [r_m^{\text{seq}}]$, we come up with the following upper-bound

$$\begin{aligned} \mathcal{V}_{\text{II}} - \mathcal{V}_{\text{III}}^{\text{seq}} &\leq \sum_{m=1}^M \mathbb{E} [\log (1 + \gamma_{[K:K],m} P_m^{\text{wf}})] - \Pr^{\text{alloc}} \\ &\quad \mathbb{E} [\log (1 + (\gamma_{[K:K],m} - \Delta_m^{\text{seq}}) P_m^{\text{seq,wf}})] \\ &\leq \sum_{m=1}^M \mathbb{E} [\log (1 + \gamma_{[K:K],m} P_m^{\text{wf}})] - \left(1 - \frac{1}{c_b}\right) \\ &\quad \mathbb{E} [\log (1 + (\gamma_{[K:K],m} - \Delta_m^{\text{seq}}) P_m^{\text{seq,wf}})] \quad (30) \\ &= U_b \quad (31) \end{aligned}$$

where P_m^{wf} and $P_m^{\text{seq,wf}}$ are the results of the water-filling of power P_{tot} over the channel gains $\gamma_{[K:K],m}$'s and $(\gamma_{[K:K],m} - \Delta_m^{\text{seq}})$'s, respectively. As seen from (23) and (30), the upper-bound for performance-loss for the sequential scheme is smaller than that for the contention-based scheme if $c_a = c_b$ due to the missing s in (30). This is because, the contention-based scheme suffers from collisions captured by the factor s . This fact is further verified by numerical results in Section IV.

C. Modifying the Proposed Protocol for WSRmax

The OF protocols proposed earlier in Sections III-A and III-B for maximizing sum-rate can be easily modified for solving the WSRmax problem (1). When the weights w_k are not the same for all users, the proposed OF protocol needs to solve (6) and not its special case (10). For a given value of λ and weights, the distribution of γ_{km} induces a distribution on the term in curly-brackets on the R.H.S. of Eqn (6), which we now define as $\zeta_{km}(\lambda)$. Thus

$$\zeta_{km}(\lambda) = \left\{ w_k \log \left(1 + \left(\frac{w_k}{\lambda} - \frac{1}{\gamma_{km}} \right)_+ \gamma_{km} \right) \right\}. \quad (32)$$

Eqn (6) can now be rewritten as

$$k_m^*(\lambda) = \arg \max_k \zeta_{km}(\lambda), \quad (33)$$

where by writing $k_m^*(\lambda)$, we have made the dependency on λ explicit. Comparing (33) and (10), intuitively, it seems that the BS can modify the proposed OF protocols by simply partitioning the pdf of $\zeta_{km}(\lambda)$ and then solve (33) by finding the user with the highest value of $\zeta_{km}(\lambda)$ in a distributed manner, the same way as the user with the highest γ_{km} was found in Sections III-A and III-B. This intuition is largely correct, with one difference that in the case of the WSRmax

problem, the reduction from (7) to (10) does not happen, hence every time the transmitter updates the value of λ , the proposed OF protocol needs to be re-run because the best user k_m^* from the solution of (33) may change as the value of λ is updated. Hence the best user for each tone is found N_{iter} times until the power constraint is satisfied.

In terms of the operation of the protocols, along with the value of p , the BS also needs to broadcast λ in the *Initiate Feedback* message shown in Figure 2. Further, since the distribution of $\zeta_{km}(\lambda)$ depends on w_k 's, it is different for different users and the users are no longer i.i.d. In order to find the user with the highest value of $k_m^*(\lambda)$ (to solve (33)) in a distributed manner, the thresholds used by all users should be identical i.e. $T_i, \forall i$, should be the same for all users. Hence, all users must compute the thresholds using the same weight value, which also should be a part of the *Initiate Feedback* message. One choice for this common weight value is $\max_k w_k$. Since the pdf of $\zeta_{km}(\lambda)$ is partitioned by the thresholds T_i 's, for the WSRmax case, the lower threshold T_i of a feedback slot i , in which the first successful feedback message is received gives an estimate of $\zeta_{km}^*(\lambda) \triangleq \max_k \zeta_{km}(\lambda)$. This estimate is denoted by $\tilde{\zeta}_{km}^*(\lambda)$. Unlike the sum-rate case, in which the lower threshold T_i serves as an estimate of the channel-gain value $\tilde{\gamma}_{km}^*$, in the WSRmax case, $\tilde{\gamma}_{km}^*$ can be found using $\tilde{\zeta}_{km}^*(\lambda)$. When $\zeta_{km}(\lambda)$ is non-zero, (32) can be rewritten as

$$\zeta_{km}(\lambda) = w_k \log \left(\frac{w_k \gamma_{km}}{\lambda} \right), \quad (34)$$

and so an estimate of channel-gain value $\tilde{\gamma}_{km}^*$ can be obtained from the estimate $\tilde{\zeta}_{km}^*$ as

$$\tilde{\gamma}_{km}^* = e^{\frac{\tilde{\zeta}_{km}^*}{w_k}} \lambda / w_k. \quad (35)$$

Using (32), the power-consumption on tone m can then be obtained as

$$p_{km} = e^{\frac{\tilde{\zeta}_{km}^*}{w_k} - 1} / \tilde{\gamma}_{km}^*. \quad (36)$$

Using this value of power-consumption (36), the BS iterates over the value of λ , until the total power-constraint is satisfied. N_{iter} denotes the number of iterations.

The fact that the distribution of $\zeta_{km}(\lambda)$ is user-dependent does not change the operation of the OF protocols, because the protocol works just as described in Sections III-A and III-B once the threshold values are computed and fixed. The analysis of the protocols, however, does become a hard task. The probability $\text{Prob}(T_i \leq \zeta_{km}(\lambda) < T_{i-1})$ for a user on a tone in a feedback slot is no longer equal to the same value p and is different for different k and i . Although the series of steps used for analysis of the proposed OF protocols in the sum-rate case can be applied for analyzing the WSRmax case as well, the resulting expressions are complicated. The feedback-overhead for both the contention-based and sequential OF protocols essentially increases linearly with N_{iter} as compared to that for the sum-rate, shown numerically in Section IV. The feedback overhead of the contention-based scheme can be expressed analytically because it is simply $N_{\text{iter}} \mathcal{F}_{\text{III}}^{\text{con}}$, where $\mathcal{F}_{\text{III}}^{\text{con}}$ is given in (17). The same does not hold for the sequential scheme because the feedback overhead depends on the compression rate, $H(p)$, and p is no longer a constant. The complexity of the proposed OF protocols, C_{III} , is also scaled by N_{iter} , which is a constant, so it remains to be constant. The performance gap is hard to derive analytically for the reasons mentioned above, and will be studied numerically in the next section.

IV. NUMERICAL RESULTS

This section presents numerical results to verify the analytical expressions derived for the two versions of Scheme-III: i) contention-based (referred to as Scheme-III_a) and ii) sequential (referred to as Scheme-III_b) that were proposed in Sections III-A and III-B, respectively, to solve the sum-rate maximization problem. For the WSRmax problem, numerical results are used to evaluate the performance of the proposed OF protocols as described in Section III-C.

The channel fading distribution for each user on each tone is i.i.d. Rayleigh fading with mean channel-gain value $\bar{\gamma}_{km} = 10, \forall k, \forall m$. Each point in the plots is obtained as an average over 200 independent channel realizations. Plots are obtained for $M = 2048$. This value is representative of the actual value used in current deployments of WiMAX (IEEE 802.16) systems. Scheme-II, as well those proposed in [19], [20] assume perfect channel knowledge i.e. channel gain values are fed back as a real number, thus using B_{real} bits. The value of B_{real} is chosen to ensure a fair comparison. For the proposed OF protocols, the pdf of channel-gain is divided into N equal probability intervals. This is the same as quantizing the channel gain values, which are real numbers, using N discrete levels. The number of bits required for this is $\log_2(N)$. Hence, the proposed OF protocols provide channel-gain value to the BS roughly⁵ with an accuracy of $\log_2(N)$ bits. So, we set $B_{\text{real}} = \log_2(N)$ for a fair comparison of the schemes. The complexity of the proposed schemes and the schemes in [19], [20] is $O(1)$, whereas that for Scheme-II is $O(MK)$. For all plots, $s = 0.9$, $P_{\text{tot}} = 20$ dB and the values of c_a and c_b are set to 3.

In Figure 3, we plot the performance in terms of sum-rate as a function of the number of users K for Scheme-II, III_a and III_b. Also plotted are the performance of the schemes proposed in [19]⁶, [20]. Since $M = 2048 \gg 8$, the performance of Scheme-II is the same as that of the optimal scheme, Scheme-I [5] and hence serves as an upper-bound for performance. Schemes III_a and III_b have a worst case performance loss of 4% and 3% with respect to the optimal value respectively. The contention-based scheme has a higher performance-loss due to the collisions. The performance of the scheme in [19] first increases with increasing the number of mini-slots, however, quickly saturates. For simulating this scheme [19], an equal integer number of users were placed in each group, except the last group which had the remaining number. This is unlike the setup used for simulations in [19], where the number of users are chosen to be an integer multiple of the number of minislots. The unequal group size explains the slight bumps in the curves for this scheme [19] with increasing total number of users K . For simulating the scheme in [20], the value of the number of clusters for grouping the M tones was chosen to be 10. The performance of the scheme in [20] was found to change negligibly for different

⁵Since the intervals between the thresholds T_i and T_{i-1} are equiprobable and not equal-length, the quantization of channel-gain value is not done uniformly by a standard uniform quantizer. However, since N is large, the quantization is roughly uniform.

⁶Since the scheme in [19] was proposed for a single carrier-system, for the OFDMA case under study, the same feedback protocol is repeated M times, once for each tone.

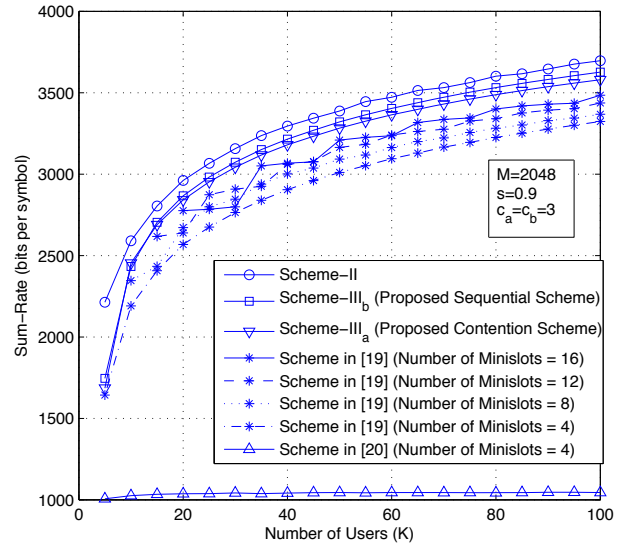
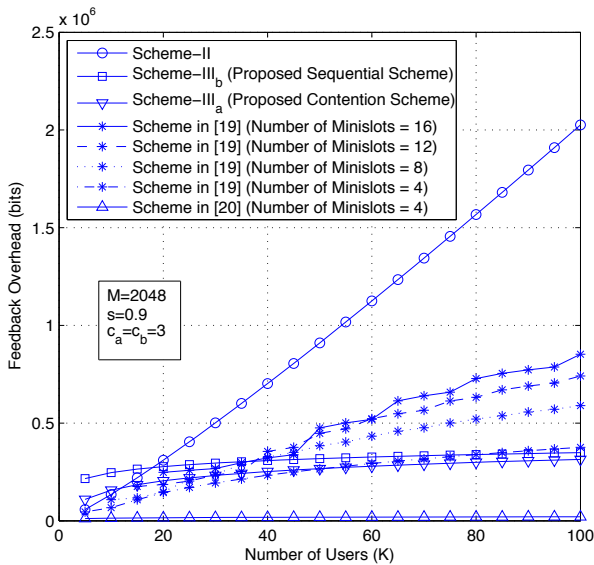
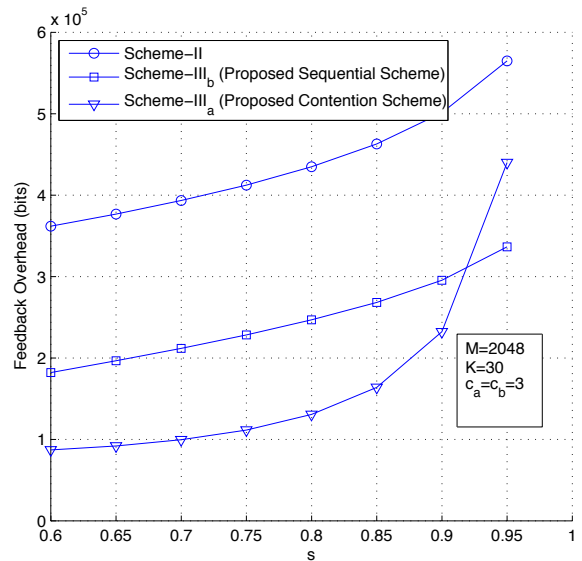
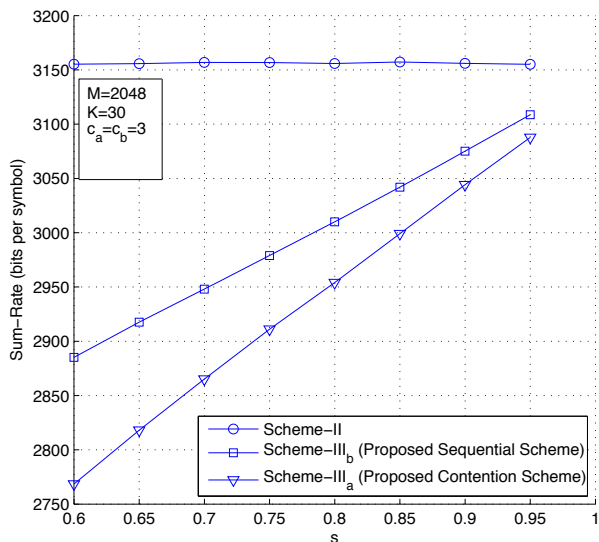
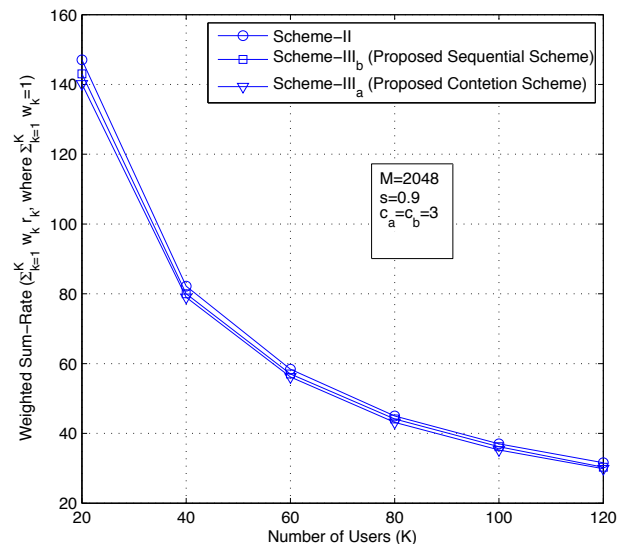


Fig. 3. Sum-Rate (\mathcal{V}) v/s K for $M=2048$

values of the number of minislots, hence is shown for one number (4) only. Increasing the number of minislots for the schemes in [19], [20] is analogous to increasing c_a , c_b or s in the proposed schemes as it trades feedback overhead with performance as seen in Figure 4. In Figure 4 the feedback overhead for the schemes is plotted for $M = 2048$. The growth rate of feedback overhead for the sequential scheme is higher (an extra factor of $\log M$ as seen from (17) and (28)) than that for the contention-based scheme. The feedback overhead for the scheme in [19] is calculated as M times ($\#$ Minislots) $\log_2 \left(\frac{K}{\# \text{ Minislots}} \right) B_{\text{real}}$, and that for the scheme in [20] is calculated as ($\#$ Clusters) ($\#$ Minislots) $\log_2 K + MB_{\text{real}}$. The extra term MB_{real} is due to second phase of feedback employed by the scheme in [20], in which the actual channel gain values of the selected users is solicited by the BS.

From Figures 3-4, it can be seen that the proposed schemes comes very close to the optimal solution in terms of performance, using much fewer feedback bits. For typical values of $K = 30$, the feedback overhead can be reduced by about 1.7 to 2 times and can be as high as 7 times for higher values $K = 100$. Additionally, although the scheme in [19] comes pretty close to the proposed schemes in performance, this happens at the cost of higher feedback overhead. The scheme in [20] requires minimal feedback overhead, however has bad performance. The scheme in [20] is perceived as a good candidate when the feedback budget is very scarce, for e.g. when the channel coherence time is very short.

In Figures 5 and 6, we study the effect of the choice of design-parameter s on sum-rate and feedback overhead for the two proposed schemes for a typical user size $K = 30$. As discussed in Section III and seen in Figure 5, the performance of both the proposed schemes increases and gets closer to the optimal value when s gets closer to 1. This, however, comes at the cost of a higher feedback overhead shown in Figure 6. The growth of feedback overhead for the contention-based protocol

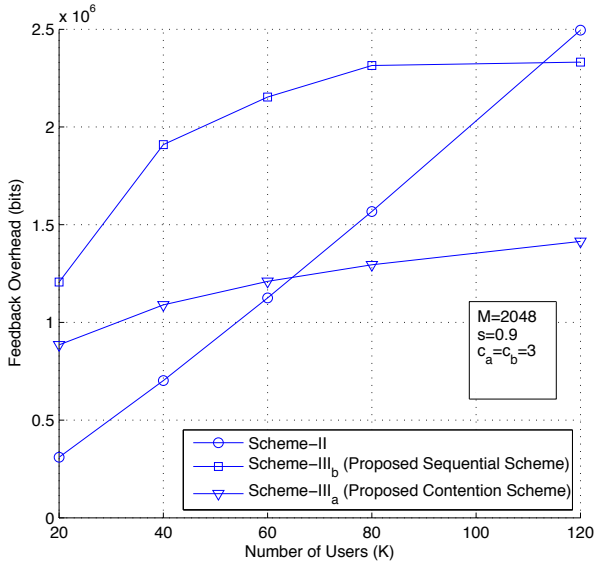
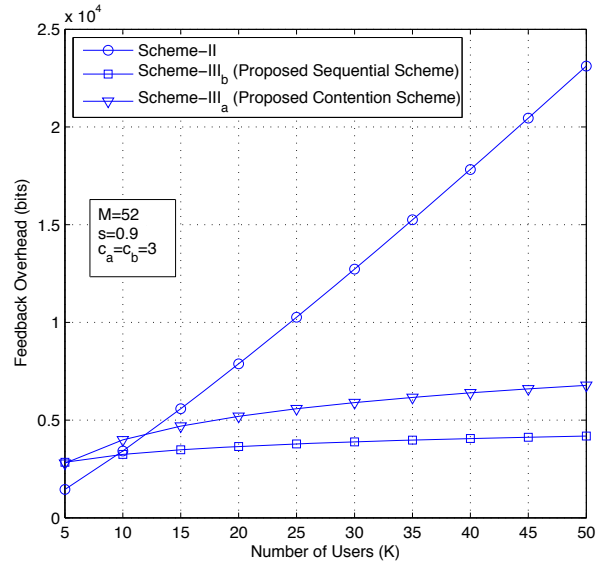
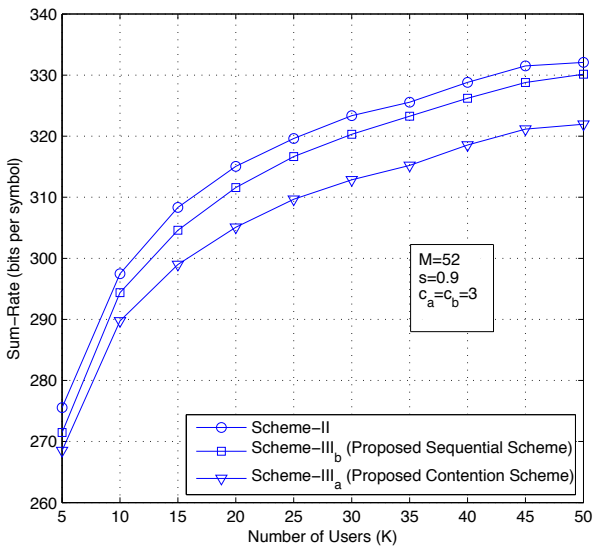
Fig. 4. Feedback Overhead (\mathcal{F}) v/s K for $M=2048$ Fig. 6. Feedback Overhead (\mathcal{F}) v/s s for $M=2048$ Fig. 5. Sum-Rate (\mathcal{V}) v/s s for $M=2048$ Fig. 7. Weighted-Sum-Rate v/s K for $M=2048$

is more steep w.r.t. s as compared to that for the sequential protocol. Hence, by varying the design parameter s , a system designer can trade-off feedback overhead with performance.

Figure 7 plots the performance of the proposed protocols in terms of weighted-sum-rate and Figure 8 shows the corresponding feedback overhead in bits. The weights for the users are generated uniformly randomly between 0 and 1, independent of each other and then normalized to sum up to 1. The number of iterations N_{iter} for the proposed protocols was chosen for an accuracy of 98% i.e. $\frac{|P_{\text{tot}} - P_{\text{used}}|}{P_{\text{tot}}} < 0.02$. As seen in Figure 7, the proposed OF protocols suffer negligible loss in performance as compared to the optimal. The weighted-sum-rate decreases with K because the downlink rate can at best grow as $\log \log K$ with the number of users K . In terms of feedback overhead, due to the additional scaling factor of

N_{iter} for the proposed OF protocols, the number of users above which the proposed schemes beat Scheme-II increases as seen in Figure 8.

An interesting observation can be made by comparing the performance of the proposed schemes for a smaller value of $M = 52$ as seen in Figures 9-10. This value is representative of the actual value used in current deployments of WiFi (IEEE 802.11 a/g) systems. Although, the growth rate of feedback overhead for the sequential scheme is higher, for the given values of parameters, the feedback overhead of the sequential scheme is smaller than that for the contention-based scheme. Hence, when M is smaller, the sequential scheme has both, better performance and lesser feedback overhead as compared to the contention-based scheme.


 Fig. 8. Feedback Overhead (\mathcal{F}) v/s K for $M=2048$

 Fig. 10. Feedback Overhead (\mathcal{F}) v/s K for $M=52$

 Fig. 9. Sum-Rate (\mathcal{V}) v/s K for $M=52$

V. CONCLUSION

To achieve high network performance, OFDMA resource allocation requires considerable feedback-overhead and has high computation complexity. Two main contributions of this paper are to exploit an opportunistic feedback protocol and a distributed computation to solve the Weighted Sum Rate Maximization (WSRmax) problem to reduce both feedback overhead and computational complexity. We propose two algorithms that exhibit constant-complexity $O(1)$ and their feedback overhead is shown to grow as $c_a M \log K$ and $c_b M \log M \log K$ bits, respectively, where M is the number of tones, K is the number of users, and c_a and c_b are constants. As proved analytically and verified through numerical results, these quantities represent a large reduction as compared to the schemes in the literature, yet the algorithms perform within a

TABLE I
SUMMARY OF THE PERFORMANCE OF VARIOUS SCHEMES

Scheme	Performance \mathcal{V}	Complexity \mathcal{C}	Overhead \mathcal{F}
Scheme-I	$\mathcal{V}_I = \mathcal{V}^{opt}$	$O(K^M)$	KMB_{real}
Scheme-II	$\mathcal{V}_{II} \xrightarrow{M \rightarrow \infty} \mathcal{V}_I$	$O(KM)$	KMB_{real}
Proposed-III _a	$\mathcal{V}_{III_a} > \mathcal{V}_{II} - U_a$	$O(1)$	$c_a M \log(K)$
Proposed-III _b	$\mathcal{V}_{III_b} > \mathcal{V}_{II} - U_b$	$O(1)$	$c_b M \log(M) \log(K)$

small loss to the optimal solution. A summary of the results in this paper is given in Table I.

REFERENCES

- [1] A. K. F. Khattab and K. M. F. Elsayed, "Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks", in Proc. *the 2006 International Symposium on World of Wireless, Mobile and Multimedia Networks*, pp. 279-288, Buffalo, NY, June 2006.
- [2] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation", *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1747-1758, Oct. 1999.
- [3] Z. Han and K. J. R. Liu, *Resource allocation for wireless networks: basics, techniques, and applications*. Cambridge University Press, 2008.
- [4] Z. Han, Z. Ji, and K. J. R. Liu, "A referee-based distributed scheme of resource competition game in multi-cell multi-user OFDMA networks", *IEEE J. Select. Areas Commun.*, vol.53, no.10, pp. 1079-1091, Aug. 2007.
- [5] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems", in proceedings of *IEEE ISIT*, pp. 1394-1398, July 2006.
- [6] D. Gesbert and M. S. Alouni, "How much feedback is multi-user diversity really worth?", in proceedings of *IEEE ICC*, vol. 1, pp. 234-238, June 2004.
- [7] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Salzer, "Efficient metrics for scheduling in MIMO broadcast channels with limited feedback", in Proc. *IEEE ICASSP*, vol. 3, pp. III-109-III-112, Apr. 2007.
- [8] D. J. Love, "Duplex distortion models for limited feedback MIMO communication", *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 766-774, Feb. 2006.
- [9] B. C. Banister and J. R. Zeidler, "Feedback assisted transmission subspace tracking for MIMO systems", *IEEE J. Select. Areas Commun.*, vol. 21, no. 3, pp. 452-463, Apr. 2003.
- [10] B. C. Banister and J. R. Zeidler, "A simple gradient sign algorithm for transmit antenna weight adaption with feedback", *IEEE Trans. Signal Processing*, vol. 51, no. 5, pp. 1156-1171, May 2003.

- [11] R. Agarwal, C. S. Hwang, and J. M. Cioffi, "Opportunistic feedback protocol for achieving sum-capacity of the MIMO broadcast channel", in proceedings of *IEEE VTC Fall*, pp. 606-610, Oct. 2007.
- [12] D. J. Love and R. W. Heath, "OFDM power loading using limited feedback", *IEEE Trans. Veh. Technol.*, vol. 54, no. 5, pp. 1773-1780, Sep. 2005.
- [13] V. K. N. Lau and T. Wu, "Optimal transmission and limited feedback design for OFDM/MIMO systems in frequency selective block fading channels", *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1569-1573, May 2007.
- [14] P. Svedman, S. K. Wilson, L. Cimini, and B. Ottersten, "Opportunistic beamforming and scheduling for OFDMA systems", *IEEE Trans. Commun.*, vol. 55, no. 5, pp. 941-952, May 2007.
- [15] S. Sanayei, A. Nosratinia, and N. Aldahir, "Opportunistic dynamic sub-channel allocation in multiuser OFDM networks with limited feedback", in proceedings of *IEEE ITW*, pp. 182-186, Oct. 2004.
- [16] I. C. Wong and B. L. Evans, "Optimal OFDMA subcarrier, rate, and power allocation for ergodic rates maximization with imperfect channel knowledge", in proceedings of *IEEE ICASSP*, vol. 3, pp. III-89-III-92, Apr. 2007.
- [17] R. Wang and V. K. N. Lau, "Cross layer design of downlink multi-antenna OFDMA systems with imperfect CSIT for slow fading channels", *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2417-2421, July 2007.
- [18] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks", in proceedings of *IEEE INFOCOM*, vol. 3, pp. 1662-1672, Mar. 2004.
- [19] S. Patil and G. de Veciana, "Feedback and opportunistic scheduling in wireless networks", *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4227-4238, Dec. 2007.
- [20] T. Tang, R. W. Heath, Jr., S. Cho, and S. Yun, "Opportunistic Feedback in Clustered OFDM Systems", in proceedings of *International Symposium on Wireless personal Multimedia Communications*, San Diego, CA, Sep. 2006.
- [21] J. Chen, R. A. Berry, and M. L. Honig, "Performance of limited feedback schemes for downlink OFDMA with finite coherence time", in Proc. *IEEE ISIT*, Nice, France, June 2007.
- [22] C. S. Hwang, K. Seong, and J. Cioffi, "Opportunistic p-persistent CSMA in wireless networks", in proceedings of *IEEE ICC*, vol. 1, pp. 183-188, June 2006.
- [23] R. Agarwal, V. Abhishek, and J. M. Cioffi, "Opportunistic feedback in OFDMA downlink", in Proc. *IEEE GLOBECOM*, pp. 3220-3224, Nov. 2007.
- [24] R. Agarwal, R. Vannithamby, and J. M. Cioffi, "Optimal Allocation of Feedback Bits for Downlink OFDMA Systems", in Proc. *IEEE ISIT*, Toronto, Canada, July 2008.
- [25] W. Yu and R. Lui, "Dual methods for non-convex spectrum optimization of multi-carrier systems", *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310-1322, July 2006.
- [26] A. Goldsmith, *Wireless communications*, Cambridge University Press, 2005.
- [27] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
- [28] H. A. David, *Order statistics*, John Wiley and Sons, Inc., 2nd ed., 1980.
- [29] T. Cover and J. Thomas, *Elements of information theory*, Wiley & Sons, Inc., 1991.



Vinay R. Majjigi completed his BS in Electrical Engineering at Case Western Reserve University, and his MS at Stanford University. He has worked in industry for two years at Boeing and Northrop Grumman. His academic interests include communications, signal processing, circuit design, and device physics. He is currently working on his PhD at Stanford under the broad area of wireless communications. On his spare time, he enjoys learning the guitar and piano, cycling, and photography.



Zhu Han (S'01-M'04) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. Currently, he is an Assistant Professor in Electrical and Computer Engineering Department at Boise State University, Idaho, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, wireless multimedia, and security. Dr. Han is the MAC Symposium vice chair of IEEE WCNC, 2008. Dr. Han is the Guest Editor for Special Issue on Fairness of Radio Resource Management Techniques in Wireless Networks, EURASIP Journal on Wireless Communications and Networking, and Special Issue on Game Theory, EURASIP Journal on Advances in Signal Processing. Dr. Han is a member of the TPC for the IEEE ICC, VTC, CCNC, WCNC and Globecom.



Rath Vannithamby received his B.S., M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Toronto, Ontario, Canada, in 1994, 1996 and 2001 respectively. He is currently a Senior Research Scientist, Manager in Corporate Technology Group at Intel Corporation, Hillsboro, Oregon, USA and leads the MAC and signaling layer standardization of next generation WiMAX system. He has published over 30 papers, and has over 50 patents pending. He has served on TPC for ICC, Globecom, VTC, and WCNC and as a guest editor for EURASIP Journal of Wireless Communications and Networking special issue on Radio Resource Management for 3G+ Systems. His current research interests are in the area of Radio Resource Management techniques, Cross-layer design and MAC/Signaling Layer Protocols for high-speed wireless access networks using OFDMA technologies including 4G and IEEE 802.16.



John M. Cioffi - BSEE, 1978, Illinois; PhDEE, 1984, Stanford; Bell Laboratories, 1978-1984; IBM Research, 1984-1986; EE Prof., Stanford, 1986-present. Cioffi founded Amati Com. Corp in 1991 (purchased by TI in 1997) and was officer/director from 1991-1997. He currently is on the Board of Directors of ASSIA (Chairman), Afond, Teranetics, and ClariPhy. He is on the advisory boards of Focus Ventures, Portview Ventures, Wavion, Quantenna, and Amicus. Cioffi's specific interests are in the area of high-performance digital transmission. Various Awards: International Marconi Fellow (2006), Holder of Hitachi America Professorship in Electrical Engineering at Stanford (2002); Member, National Academy of Engineering (2001); IEEE Kobayashi Medal (2001); IEEE Millennium Medal (2000); IEEE Fellow (1996); IEE JJ Tomson Medal (2000); 1999 U. of Illinois Outstanding Alumnus, and others. Cioffi has published over 250 papers and holds over 80 patents, of which many are heavily licensed including key necessary patents for the international standards in ADSL, VDSL, DSM, and WiMAX.



Rajiv Agarwal completed his B.Tech. in Electrical Engineering from I.I.T. Kanpur, and his M.S. from Stanford University in 2003 and 2005 respectively. He has worked in the industry for one year at Marvell Semiconductors, Santa Clara, CA. His academic interests include wireless communications, information theory and machine learning. He is currently working on his PhD at Stanford. Rajiv has over 13 technical publications in IEEE conferences and journals. He has served as session chairs at the IEEE VTC, Spring 2007 and IEEE WCNC 2008.

He was a member of the TPC for the IEEE WCNC 2008.