Queue-based Sub-carrier Grouping For Feedback Reduction In OFDMA Systems

Harish Ganapathy and Constantine Caramanis
† Department of Electrical and Computer Engineering
The University of Texas, Austin
Austin, TX 78712, USA
E-mail: harishg@utexas.edu,caramanis@mail.utexas.edu

Abstract*—Sub-carrier grouping is a popular feedback reduction approach for orthogonal-frequency-division multiple-access (OFDMA) systems that has been adopted into fourth-generation standards such as 3GPP Long Term Evolution (LTE). Feedback reduction is motivated by the fact that the bandwidth expenditure in acquiring full information in a downlink OFDMA system scales as the product of the number of users and the number of OFDMA bands. As this is infeasible in most systems, sub-carrier grouping calls for users to report a single channel state value per pre-designated group of OFDM bands. Such an approach would reduce the amount of feedback by a factor that is equal to the size of the group, albeit at a loss in throughput. In this paper, we propose a throughput-optimal joint sub-carrier grouping and data scheduling policy that makes decisions based on queue-lengths and channel states in each time slot. The feedback allocation or sub-carrier grouping policy, inspired by the current approach in LTE, operates under a total feedback budget and must periodically decide a sub-carrier grouping size for each user that obeys this resource constraint. However, as we show, the optimal allocation algorithm has complexity that, in general, scales exponentially in the number of users. Thus, we turn our attention to the important issue of computational efficiency and propose a greedy algorithm that allocates full feedback bandwidth to a constant-sized subset of users based on the network state. We evaluate the performance of this simple approach through extensive numerical experiments. We show that under asymmetric arrival rate settings, our greedy algorithm is within 10% of the optimal (throughput-wise) when consuming only 25% of the full feedback bandwidth while paying only a logarithmic (in the number of users) price in control overhead.

I. Introduction

Over the last decade, there has been an ever-increasing demand for data-rate in wireless systems coupled with a growing scarcity for spectrum. To support these throughput demands, network service providers are on the constant search for new transmission technologies to utilize available spectrum in as efficient a manner as possible. At the physical layer, a key fundamental enabler of these new technologies such as multiple antennas is the presence of feedback channels. Feedback channels provide channel state information (CSI) to the transmitter thereby allowing it to adapt its transmission strategy to the fluctuating channel and realize the gains promised by the new methods. That said, feedback channels do consume valuable bandwidth and hence, it is important to ensure that the net gains post feedback remain justifiable. To understand the typical feedback requirements of an orthogonal-frequency-division multiple-access (OFDMA) system, let us consider an LTE-inspired example recently provided by Ouyang et al. [1].

In LTE, the smallest unit of bandwidth that can be assigned to a user for data transmission is called a resource block, which is essentially a group of OFDM sub-carriers. If we consider a 10MHz LTE system bandwidth and set $L = 50$ to be the number of resource blocks shared by $K = 50$ users equipped with standard 4-bit modulation/coding tables at the mobiles, we have a total feedback bandwidth of $4KL = 4 \times 50 \times 50 = 10$kb per sub-frame [3]. Given a typical uplink data rate of 48kb per sub-frame, this consumes 20% of the uplink capacity, which is clearly quite significant. Furthermore, it is foreseeable that future multiple-antenna systems will allow for more resolution, potentially as small as one sub-carrier per resource block, i.e., $L = 1024$ for a 10MHz system with 1024 OFDM sub-carriers. Thus, the feedback bandwidth could be as large as $4rKL = 4 \times 2 \times 50 \times 1024 \approx 200$kb, where $r = 2$ accounts for 2×2 multiple-input-multiple-output (MIMO) transmissions that are already part of the LTE standard. In this situation, if the corresponding uplink capacity does not exhibit super-linear scaling in the feedback resolution, then feedback bandwidth consumption remains an important bottleneck.

A. Prior work on feedback design

To alleviate this feedback requirement, various feedback reduction schemes have been proposed in the literature of which one simple, yet effective, scheme is to group together multiple sub-carriers (or resource blocks) and provide a single CSI report to the base station for the entire group. This is called *sub-band-level feedback* in LTE. The other methods considered in LTE are *UE selected sub-band feedback* where the mobile selects its best few sub-bands and reports the average CSI value on these channels and *wideband feedback* where the mobile reports one value for the entire bandwidth. We focus on the first method through the remainder of this paper. The merits of sub-band feedback and variants of it have been studied in detail over recent years [4]–[8]. Donthi and Mehta [4] derive in closed-form the throughput of the sub-carrier grouping technique under various schedulers such as round robin, proportional fairness and max-sum-rate. Hybrid schemes have been considered that combine sub-carrier grouping and channel thresholing. Here, the CSI is reported for the group if its quality exceeds a pre-specified threshold [6]–[8]. Jorsweick et al. [10] consider an uplink MIMO-OFDM system and propose a scheme that reports a single covariance matrix for a chunk of carriers.

The aforementioned literature focuses on metrics such as ergodic sum-rate that are more applicable to full-buffer (sat-

urated) systems. In contrast, there has been relatively little research on the impact of limited feedback on queueing (unsaturated) systems [1], [11], [14] where feedback allocations are made as a function of *the queue sizes in addition to the channel strengths*. The feedback allocations typically respect a total budget constraint in each scheduling slot that has been modelled in a variety of ways. Our earlier work [14] introduces a limited feedback model for uplink systems where the base station is constrained in the number of bits that form the feedback packet that it broadcasts to the users. Computationally-efficient algorithms are presented that compute the optimal (or near-optimal) feedback partitioning across users as a function of the channel and queue state as well as other network parameters. A key assumption that is made here [14] is that the data scheduling decisions are made on a slower time-scale, as is the case in *(semi-)persistent* scheduling that is part of the LTE standard. As a result, user assignments are assumed fixed through the course of feedback optimization. Ouyang et al. [1] consider the downlink of an OFDMA network with a limited feedback model where the base station is able to acquire channel state information on a restricted number of frequency bands. In particular, each user is instructed to report CSI for at most F_i bands such that $\sum_i F_i$ does not exceed the total feedback budget. Following CSI acquisition, MaxWeight scheduling [12] is performed. The authors develop an order-wise throughput-optimal feedback allocation policy, which essentially selects a feedback partition $\{F_i\}$ in each time slot as a function of the queue lengths and channel strengths.

B. Joint feedback allocation and data scheduling

In this paper, we propose a protocol that jointly allocates feedback and data transmission resources in a downlink OFDMA system with L sub-bands/sub-carriers and K users. The protocol operates on two time-scales: MaxWeight data scheduling occurs on the faster time-scale (e.g., 1ms in LTE). Feedback allocation is done less often on the time scale of large-scale fading so that the additional overhead required to enable such an optimization is justifiable from the perspective of a network provider. The feedback allocation process is inspired by the approach in LTE [20] and determines the optimal sub-carrier grouping factor for each user based on queues and channels such that the total feedback bandwidth across all users does not exceed B bits where $B \leq KL \log_2 M$; M is the size of the standard modulation/coding scheme (MCS) table that resides at each mobile. Note that a grouping factor of L would result in a feedback bandwidth of $K \log_2 M$ bits while a factor of one would correspond to the full feedback bandwidth of $KL \log_2 M$ bits.

Our work clearly differs from Ouyang et al. [1] in that we consider an alternate feedback reduction policy that is motivated by the standards. In a broad sense, we build on our previous work by considering data scheduling on the faster time-scale. We continue to pursue the two time-scale resource allocation approach that we introduced earlier [14] in the context of limited feedback and queueing systems. We re-iterate that in the absence of such a feature, the optimization of the feedback resource might introduce unacceptable control overhead over-on-top the feedback bandwidth, which in itself might be perceived as control overhead.

C. Main contributions

The main contributions of this paper are as follows:

1) We propose a throughput-optimal joint sub-carrier grouping and data scheduling policy that operates on two time-scales.

Once we identify the correct feedback allocation problem to solve on the slower-time scale, we turn our attention to the issue of computational complexity.

2) As we will see, the complexity of determining the optimal allocation grows exponentially in the number of users. However, we identify simple CSI reporting mechanisms, which induce specific structure that allow us to bound and hence optimize the objective function.
3) We propose a simple linear-time greedy heuristic that essentially chooses the "best-expected" subset of users and allocates feedback resources *only to these users*. When the number of OFDMA sub-carriers is $L = 1$, and the total feedback budget is set to $B = cL \log_2 M,\ c \in \{1, 2, \ldots, K\}$, our greedy algorithm essentially selects the best-expected c users for full feedback prior to MaxWeight scheduling, an idea that was first proposed by Gopalan et al. [11]. Thus, our work can be seen as a generalization of their work towards wideband systems.
4) Through extensive numerical experiments, we show that the greedy algorithm performs close to the full CSI algorithm (and hence the optimal grouping algorithm) while consuming considerably less feedback bandwidth. In particular, we identify asymmetric load settings where the algorithm achieves within 10% of the maximum throughput with a feedback bandwidth of only 25%, parameterized by $c = 8$ and $K = 32$ users in the system. In addition, a favourable property of the proposed algorithm is that the amount of control overhead scales only logarithmically in the number of users.

The rest of this paper is organized as follows. In Section II, we introduce the network model with emphasis on the role of the feedback. In Section III, we discuss throughput-optimal online scheduling policies with sub-carrier grouping. We study the structure of the optimal feedback allocation policy under specific CSI reporting mechanisms in Section IV. A fast greedy feedback allocation scheme is proposed in Section V. The performance of the algorithm is evaluated through numerical experiments in Section VI. We conclude the paper with some remarks in Section VII.

II. System model

We consider the downlink of a frequency-division-duplex (FDD), OFDMA system with L sub-carriers/sub-bands and K users that operates in slotted-time. The network model is described below:

Channel State: The maximum supportable rate for user i on sub-band j at time t is given by $X_{ij}(t)$. We assume that $X_{ij}(t)$ is ergodic and comes from a finite set $\mathcal{M} = \{r_1, \ldots, r_M\}$ that essentially models an MCS table, now a regular feature at the mobile end. The cumulative distribution function for rate $X_{ij}(t)$ is given by

$$\Pr\left(X_{ij}(t) = r_m\right) = \rho_{mi}\left(\alpha_i\left(t\right)\right) \tag{1}$$

where $\alpha_i(t)$ denotes a large-scale fading gain that is dependent on user position. Users change positions once every T_{LS} slots where $T_{LS} \in \{1,2,3,\ldots\}$ denotes the large-scale fading coherence time. For ease of notation, we introduce a counter $\bar{t} = \lfloor \frac{t}{T_{LS}} \rfloor T_{LS}$ to keep track of the slower large-scale fading time-scale, i.e., $\rho_{mi}(\alpha_i(t)) = \rho_{mi}(\alpha_i(\bar{t}))$, $\forall t$. For convenience, we also set $\rho_{mi} = \rho_{mi}(\alpha_i(\bar{t}))$ making implicit the dependence on t and T_{LS}. Note that the large-scale coefficient is typically only distance-dependent and independent of frequency allowing us to omit the index j when representing it. We assume that the base station has perfect knowledge of $\{\alpha_i(\bar{t})\}$ and all distribution information $\{\rho_{mi}\}$.

Traffic model and network state: Each user k, $k = 1,2,\ldots,K$, has a queue of untransmitted packets with queue-length $q_k(t)$ that is maintained at the base station with associated arrival rate λ_k. The network state at time t is given by $\vec{M}(t) = \left(\{X_{ij}(t)\}, \vec{Q}(t)\right)$ where $\vec{Q}(t) = [q_1(t)\ q_2(t) \ldots q_K(t)]^T$.

Feedback model: Let the sub-carriers in our OFDMA system be indexed by $\mathcal{S} = \{1,2,\ldots,L\}$. As OFDM transmissions employ Fast Fourier Transforms to place user data on each sub-band, we assume that L is a power of two. We now define a partitioning or grouping scheme for these sub-carriers that creates a uniform partition of set $\mathcal{S}$ for a given user. The partition can potentially be different across users. More formally, given $g_i \in \{0,1,\ldots,\log_2 L\}$, we create $L2^{-g_i}$ groups or partitions $\mathcal{P}_i = \left\{\mathcal{S}_{1i}, \mathcal{S}_{2i}, \ldots, \mathcal{S}_{\frac{L}{2^{g_i}}i}\right\}$, $\bigcup_{p=1}^{\frac{L}{2^{g_i}}} \mathcal{S}_{pi} = \mathcal{S}$, $\mathcal{S}_{pi} \cap S_{qi} = \emptyset$, $\forall i$, $p \neq q$, such that

$$\mathcal{S}_{pi} = \{(p-1)2^{g_i}+1, (p-1)2^{g_i}+2, \ldots, p2^{g_i}\},\ p = 1,\ldots,\frac{L}{2^{g_i}}. \tag{2}$$

For example, for $L = 8$ and $g_i = 1$, we have $\mathcal{S}_{1i} = \{1,2\}$, $\mathcal{S}_{2i} = \{3,4\}$, $\mathcal{S}_{3i} = \{5,6\}$ and $\mathcal{S}_{4i} = \{7,8\}$. In general, different grouping factors for each user are permitted. Let $\vec{g} = [g_1\ g_2 \ldots g_K]^T$ parameterize a system-wide partitioning $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K\}$. At the beginning of each large-scale fading instant, the base station decides $\vec{g}^*(\bar{t})$ based on $M(\bar{t})$ and some suitable cost criterion. We refer to this process as *feedback allocation*. The total feedback bandwidth consumed by allocation $\vec{g}$ is $(\sum_{k=1}^K 2^{-g_k}) L \log_2 M$. We assume that the feedback channel has a total bandwidth of B bits, which means any allocation must respect the constraint

$$L \log M \left(\sum_{k=1}^K 2^{-g_k}\right) \leq B.$$

Once the allocation decision is made at the beginning of each large-scale fading instant, the decision is then communicated to the mobiles incurring an overhead of at most $K \log_2 \log_2 L$ bits per large-scale coherence time. Mobile i adopts the allocation decision $\vec{g}^*$ until the next large-scale fading instant (i.e., through slots $\bar{t} \leq t < \bar{t}+1$) and in turn reports $\frac{L}{2^{g_i^*}}$ "effective" CSI values, one for each group in $\mathcal{P}_i^* = \left\{\mathcal{S}_{1i}, \mathcal{S}_{2i}, \ldots, \mathcal{S}_{\frac{L}{2^{g_i^*}}i}\right\}$ according to the following rule

$$X_{ip}^{eff}(t) = f\left(\{X_{ij}\}_{j \in \mathcal{S}_{pi}}\right),\ p = 1,\ldots,\frac{L}{2^{g_i^*}}. \tag{3}$$

Herein, we refer to $f(\cdot)$ as the *quantization function*. Candidate functions that are of interest and have been studied in the past include $f(x_1,\ldots,x_n) = \min_i x_i$, $f(x_1,\ldots,x_n) = \frac{1}{n}\sum_i x_i$ and $f(x_1,\ldots,x_n) = \max_{x \in \mathcal{M}} x \sum_{i=1}^n \mathbb{I}(x > x_i)$. These three functions basically correspond to reporting the minimum supportable rate if the user is scheduled on any sub-band in the group, the average rate across the group, and the maximum supportable rate (or "goodput" as it is sometimes called) if the user were to be scheduled on all the sub-bands in the group. Later in the paper, we analyse the first quantization function in detail. Note that the case $g_i = 0$, $\forall i$, represents the case with full feedback where each user reports CSI on all sub-bands (i.e., $X_{ip}^{eff}(t) = X_{ip}(t)$, $p = 1,2,\ldots,L$) and the total feedback bandwidth is $KL \log_2 M$. In what follows, we set $B = cL \log M$, $c \in (0,K]$ for convenience allowing us to express the feedback budget constraint concisely as $\sum_{k=1}^K 2^{-g_k} \leq c$.

In the next section, we develop a two time-scale throughput-optimal sub-carrier grouping and data allocation policy.

III. Scheduling policies under limited feedback

In this section, we develop a sub-carrier grouping protocol that when operated in conjunction with the MaxWeight data scheduling policy [12] guarantees throughput-optimality. This means that given an arrival rate vector $\vec{\lambda}$, if there exists any scheduling policy that can guarantee bounded expected queue sizes, then so can the proposed policy.

The policy operates in two-stages: feedback allocations are made on the slower time-scale, namely every T_{LS} slots as a function of the queue-state $\vec{Q}(\bar{t})$ and channel statistics. This is followed by MaxWeight scheduling during instants $t = \bar{t},\ldots,\bar{t}+1$ until the next feedback allocation cycle. We now describe the policy more formally followed by a discussion of the key features of the same.

Algorithm 1 Joint feedback allocation and data scheduling

1: **for** $t = \bar{t} \ldots \bar{t} + T_{LS} - 1$ **do**

2: *(Feedback allocation)*: The sub-carrier grouping is given as the solution to

$$\begin{aligned} \vec{g}^*(\bar{t}) = \ & \arg\max \quad \mathbb{E}\left[\textstyle\sum_j \max_i Q_i(\bar{t}) X_{ij}^{eff}(t) \,\middle|\, \vec{Q}(\bar{t})\right] \\ & \text{s.t.} \quad \textstyle\sum_{k=1}^K 2^{-g_k} \leq c \\ & \quad\ g_k \in \{0,1,\ldots,\log_2 L, \infty\}, \end{aligned} \tag{4}$$

where the expectation is computed over the channel.

3: *(MaxWeight data scheduling)*: Given $\vec{g}^*(\bar{t})$, the users are scheduled according to

$$\begin{aligned} \{W_{ij}^{LF}(t)\} = \ & \arg\max \textstyle\sum_{i,j} Q_i(\bar{t}T_{LS}) X_{ij}^{eff}(t) W_{ij}(t) \\ & \text{s.t.} \quad W_{ij}(t) \in \{0,1\}, \\ & \quad\ \textstyle\sum_i W_{ij}(t) \leq 1, \forall j. \end{aligned} \tag{5}$$

4: **end for**

An allocation $g_k = \infty$ in (4) means that user k is assigned no feedback bandwidth. Note that $X_{ij}^{eff}(t)$ in (5) is a function

of $\vec{g}^*(\bar{t})$. A subtle technical point in (4) is that since we are making decisions before viewing the exact realizations of the channel, we appeal to the ergodicity of the channel process in computing the expectation in (4). The proposed algorithm is also applicable to some uplink settings while recognizing a mere philosophical difference in that on the uplink, we have access to perfect information at the base station but are unable to communicate this information to the user due to feedback constraints. The scenario considered by Jorsweick et al. [10] is illustrative. Here, the optimization problem in (4) would essentially decide the group of sub-bands that use the same MIMO precoder.

There is an important difference between the rule in (5) and the more traditional form of MaxWeight given by

$$\max_{W_{ij}(t)\in\{0,1\},\sum_i W_{ij}(t)\leq 1,\forall j} \sum_{i,j} Q_i(t) X_{ij}^{eff}(t) W_{ij}(t)$$

in that we update the queue states on a slower time scale. From past works on throughput optimality in queuing systems [16], we know that queue information acquired on a slower time scale does not affect the throughput region of the system. This model is suitable for networks where the base station does not have access to the entire queue state in every scheduling slot such as when the queues reside at the radio network controller, or on the uplink. In scenarios where the queue state is indeed available to the base station in every scheduling slot, we choose to study the above policy for analytical tractability.

This brings us to our first result. The following theorem establishes the throughput-optimality of our proposed algorithm. The proof is straightforward and relies on the notion of *Static Service Split* feedback allocation policies that are introduced in earlier papers [14], [19] and are well-understood.

Theorem 1. *The two-stage feedback allocation and data scheduling protocol given by (4) and (5) is throughput-optimal.*

The quantity

$$d_F\left(\vec{Q},\vec{\alpha},\vec{g}\right) = \mathbb{E}\left[\sum_j \max_i Q_i X_{ij}^{eff} \;\middle|\; \vec{Q}\right]$$

in (4) is henceforth referred to as *expected queue-weighted drain* and plays an important role in most resource allocation problems (e.g. [1], [11]) that involve making decisions *before* we can view the channel realizations.

At the start of each large-scale fading instant $\bar{t}$, we are interested in choosing a feedback allocation vector $\vec{g}(\bar{t})$ that maximizes the expected queue-weight drain $d_F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}(\bar{t})\right)$ subject to the total feedback budget constraint. There are two immediate challenges in finding a solution to the above problem. Firstly, given an allocation, the computation of an expectation over the channel might be computationally-prohibitive for large input (number of users, sub-bands, size of channel space, etc.) sizes. Secondly, any brute-force approach to the computation of the optimal solution will require searching through an exponential number of possibilities, specifically $\mathcal{O}\left((\log_2 L+1)^K\right)$.

The remainder of this paper is devoted to addressing these two main challenges. We proceed by assuming a specific form for the quantization function $f(\cdot)$. This enables us to derive a closed-form lower bound on $d_F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}(\bar{t})\right)$, that is "seldom loose", thereby addressing the first issue. This development, in turn, allows us to propose a linear-time greedy heuristic in Section V that optimizes the lower bound and thus addresses the second issue. We evaluate the performance of the heuristic in Section VI by comparing it against the case with full feedback bandwidth.

IV. Lower bound on expected queue-weighted drain

In this section, we consider a particular quantizer function that allows us to bound the expected queue-weighted drain $d_F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}(\bar{t})\right)$ in closed-form. In particular, we consider

$$f(x_1,x_2,\ldots,x_n) = \min_{s\in\{1,2,\ldots,n\}} x_s, \tag{6}$$

which means that if a user is scheduled on a sub-band, then the base station transmits at the lowest rate that is supported across the group that the sub-band belongs to. We note that this is a conservative choice for transmission rate that essentially guarantees zero outage making it suitable for delay-intolerant applications.

To aid in the analysis, we let $F_i(j)\in\left\{1,2,\ldots,\frac{L}{2^{g_i}}\right\}$ denote the partition that sub-band $j\in\mathcal{S}$ belongs to under feedback allocation g_i. For notational convenience with the queue-length vector, we drop mention of their dependence on t. Ergodicity allows us to do the same with the channel state variables as well. We proceed by re-writing $d_F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}(\bar{t})\right)$ as follows:

$$\begin{aligned}
&d_F\left(\vec{Q},\vec{\alpha},\vec{g}\right)\\
=&\mathbb{E}\left[\max\textstyle\sum_{i,j} Q_i X_{ij}^{eff}(t) W_{ij}(t) \;\middle|\; \vec{Q}\right]\\
=&\textstyle\sum_j \mathbb{E}\left[\max\sum_i Q_i X_{ij}^{eff}(t) W_{ij}(t) \;\middle|\; \vec{Q}\right]\\
=&\textstyle\sum_j \mathbb{E}\left[\sum_i Q_i X_{ij}^{eff}(t) W_{ij}^{LF}(t) \;\middle|\; \vec{Q}\right]\\
=&\textstyle\sum_j\sum_i \Pr\left(W_{ij}^{LF}(t)=1 \;\middle|\; \vec{Q}\right)\mathbb{E}\left[Q_i X_{ij}^{eff}(t) \;\middle|\; \vec{Q}, W_{ij}^{LF}(t)=1\right]\\
=&L\textstyle\sum_i \Pr\left(W_i^{LF}(t)=1 \;\middle|\; \vec{Q}\right)\mathbb{E}\left[Q_i X_{ij}^{eff}(t) \;\middle|\; \vec{Q}, W_i^{LF}(t)=1\right].
\end{aligned} \tag{7}$$

The last step follows from the fact that for a given user, all OFDMA bands have statistically equivalent channels, which allows us to drop the dependence on j from our notation, when appropriate. We also drop mention of the conditioning on $\vec{Q}$ through most of the section as this is understood by now. Finally in (7), we assumed all users are active, i.e., $Q_k>0,\ \forall k$, to simplify notation.

To proceed with the analysis, we now observe that the event $\{Q_i X_{ij}^{eff} > \max_{k\neq i} Q_k X_{kj}^{eff}\}$ implies that user i is scheduled. Thus, we can bound the queue-weighted drain in (7) as follows

$$\begin{aligned}
&d_F\left(\vec{Q},\vec{\alpha},\vec{g}\right)\\
\geq\ & L\textstyle\sum_i \Pr\left(Q_i X_{ij}^{eff} > \max_{k\neq i} Q_k X_{kj}^{eff}\right)\mathbb{E}\left[Q_i X_{ij}^{eff}\right.\\
&\left.\;\middle|\; Q_i X_{ij}^{eff} > \max_{k\neq i} Q_k X_{kj}^{eff}\right]\\
=\ & L\textstyle\sum_i R_i^{LF}\left(\vec{Q},\vec{\alpha},\vec{g}\right),
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
R_i^F\left(\vec{Q},\vec{\alpha},\vec{g}\right) =& \Pr\left(Q_i X_{ij}^{eff} > \max_{k\neq i} Q_k X_{kj}^{eff}\right)\times\\
&\mathbb{E}\left[Q_i X_{ij}^{eff}(t) \;\middle|\; Q_i X_{ij}^{eff} > \max_{k\neq i} Q_k X_{kj}^{eff}\right]
\end{aligned}$$

is essentially a lower bound on the queue-weighted drain for user i in particular. The bound is tight, i.e., $d_F(\vec{Q},\vec{\alpha},\vec{g}) = L\sum_i R_i^F(\vec{Q},\vec{\alpha},\vec{g})$ if $\mathcal{T}(\vec{Q}) = \emptyset$ where

$$\mathcal{T}(\vec{Q}) = \left\{(x_1,\ldots,x_K) \in \mathcal{M}^K : Q_i x_m = Q_j x_n,\ i \neq j\right\}$$

represents the (sub)set of all channel states that leads to ties during data scheduling for the given queue sizes in $\vec{Q}$. For any sufficiently random arrival process with large packets and a sufficiently small channel space $|\mathcal{M}|$, we note that ties are unlikely, which means $\mathcal{T}(\vec{Q}) \approx \emptyset$. Thus, for all practical purposes, if we can compute $R_i^F(\vec{Q},\vec{\alpha},\vec{g})$ exactly for all i, we have a good estimate of the queue-weighted drain $d_F(\vec{Q},\vec{\alpha},\vec{g})$.

To proceed with the exact computation of $R_i^F(\vec{Q},\vec{\alpha},\vec{g})$, we condition on the maximum weighted-channel amongst the other $K-1$ users as follows

$$R_i^F\left(\vec{Q},\vec{\alpha},\vec{g}\right) = \sum_{m=1}^{M+1} \Pr\left(X_{ij}^{eff} > Y_i \;\middle|\; Y_i \in \mathcal{A}_m\right) \mathbb{E}\left[Q_i X_{ij}^{eff} \;\middle|\; X_{ij}^{eff} > Y_i,\ Y_i \in \mathcal{A}_m\right] \Pr\left(Y_i \in \mathcal{A}_m\right), \tag{9}$$

where $Y_i = \max_{k\neq i} \sigma_{ki} X_{kj}^{eff}$, $\sigma_{ki} = \frac{Q_k}{Q_i}$ and

$$\mathcal{A}_n = \begin{cases} [0, r_1), & n = 1 \\ [\, r_{n-1}, r_n\,), & n = 2,3,\ldots,M \\ [\, r_n, \infty\,), & n = M+1. \end{cases} \tag{10}$$

Now since $X^{eff} \in \mathcal{M}$ by definition, we have that

$$\Pr\left(X_{ij}^{eff} > Y_i \;\middle|\; Y_i \in \mathcal{A}_m\right) = \Pr\left(X_{ij}^{eff} \geq r_m \;\middle|\; Y_i \in \mathcal{A}_m\right)$$

for $m = 1,2,\ldots,M$, and $\Pr\left(X_{ij}^{eff} > Y_i | Y_i \in \mathcal{A}_{M+1}\right) = 0$. Hence, we can ignore the last term and re-write (9) as

$$R_i^F\left(\vec{Q},\vec{\alpha},\vec{g}\right) = \sum_{m=1}^{M} \Pr\left(X_{ij}^{eff} \geq r_m \;\middle|\; Y_i \in \mathcal{A}_m\right) \mathbb{E}\left[Q_i X_{ij}^{eff} \;\middle|\; X_{ij}^{eff} \geq r_m,\ Y_i \in \mathcal{A}_m\right] \Pr\left(Y_i \in \mathcal{A}_m\right). \tag{11}$$

It also follows from the boundedness of the channel space that $\Pr\left(X_{ij}^{eff} \geq r_1 \;\middle|\; Y_i \in \mathcal{A}_1\right) = 1$ and

$$\mathbb{E}\left[Q_i X_{ij}^{eff} \;\middle|\; X_{ij}^{eff} \geq r_1,\ Y_i \in \mathcal{A}_1\right] = Q_i \mathbb{E}\left[X_{ij}^{eff}\right].$$

.

In Lemma 1, we compute in closed-form the three quantities $\Pr\left(X_{ij}^{eff} \geq r_m | Y_i \in \mathcal{A}_m\right)$, $\mathbb{E}[Q_i X_{ij}^{eff} | X_{ij}^{eff} \geq r_m,\ Y_i \in \mathcal{A}_m]$ and $\Pr\left(Y_i \in \mathcal{A}_m\right)$ of interest in (11). These results are combined to yield an exact closed-form expression for $R_i^F\left(\vec{Q},\vec{\alpha},\vec{g}\right)$ in Theorem 2. All quantities are expressed as a function of the complementary cumulative distribution function (CCDF) of the users' channels; $P_{X_{ij}}(r) = \Pr\left(X_{ij} \geq r\right) = \sum_{\{n: r_n \geq r\}} \rho_{ni}$ denotes the CCDF of user i's channel, assumed available to the base station in closed-form.

Lemma 1. (a) *The conditional probability* $Pr\left(X_{ij}^{eff} \geq r_m \;\middle|\; Y_i \in \mathcal{A}_m\right)$ *is given by*

$$Pr\left(X_{ij}^{eff} \geq r_m \;\middle|\; Y_i \in \mathcal{A}_m\right) = \left[P_{X_{ij}}(r_m)\right]^{2^{g_i}}. \tag{12}$$

(b) *The probability* $Pr\left(Y_i \in \mathcal{A}_m\right)$ *is given by*

$$Pr\left(Y_i \in \mathcal{A}_m\right) = \prod_{k\neq i}\left[1 - \left(P_{X_{k1}}\left(\tfrac{r_m}{\sigma_{ki}}\right)\right)^{2^{g_k}}\right] + \prod_{k\neq i}\left[1 - \left(P_{X_{k1}}\left(\tfrac{r_{m-1}}{\sigma_{ki}}\right)\right)^{2^{g_k}}\right]. \tag{13}$$

(c) *The conditional expectation* $\mathbb{E}\left[Q_i X_{ij}^{eff} \;\middle|\; X_{ij}^{eff} \geq r_m,\ Y_i \in \mathcal{A}_m\right]$ *is given by*

$$\mathbb{E}\left[Q_i X_{ij}^{eff} \;\middle|\; X_{ij}^{eff} \geq r_m,\ Y_i \in \mathcal{A}_m\right] = \frac{Q_i}{\sum_{p=1}^{m-1}\chi_p} \sum_{n=m}^{M} r_n \chi_{in} \tag{14}$$

where $\chi_{in} = \left(P_{X_{i1}}(r_n)\right)^{2^{g_i}} - \left(P_{X_{i1}}(r_{n+1})\right)^{2^{g_i}}, n = 1,2,\ldots,M$, *and* $r_{M+1} = \infty$.

Proof: Refer to Appendix A. ∎

Consolidating the results in Lemma 1, we can obtain the desired closed-form bound on the expected queue-weighted drain for user i and hence the expected queue-weighted drain for the system.

Theorem 2. *The expected queue-weighted drain of user* i *can be bounded below as shown in (15) where* $\chi_{in} = \left(P_{X_{i1}}(r_n)\right)^{2^{g_i}} - \left(P_{X_{i1}}(r_{n+1})\right)^{2^{g_i}}, n = 1,2,\ldots,M$ *and* $r_{M+1} = \infty$. *Equality is achieved if* $\mathcal{T}(\vec{Q}(\bar{t})) = \emptyset$.

Proof: Follows from (9) and Lemma 1. ∎

Having derived a closed-form bound that we believe is tight for most practical settings, the feedback allocation problem in (4) now takes the form

$$\begin{aligned} \vec{g}(\bar{t})^* = \quad & \arg\max \quad \textstyle\sum_i R_i^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}\right) \\ & \text{s.t.} \quad \textstyle\sum_{k=1}^{K} 2^{-g_k} \leq c \\ & \quad\ g_k \in \{0,1,\ldots,\log_2 L,\infty\} \end{aligned} \tag{15}$$

where $R_i^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}\right)$ is given in Theorem 2. With the new formulation in (15), we have successfully addressed the issue of objective-function-computation raised at the end of the previous section. The objective function in (15) can be calculated using $\mathcal{O}(KM(K+M))$ operations, which is $\mathcal{O}(K^2)$ for most realistic settings, where M does not scale.

Next, we turn our attention to the second issue raised at the end of the previous section and this is the topic of computational-efficiency. In the developments that follow, we refer to $R_i^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}\right)$ as the expected queue-weighted drain for user i thereby accepting it as an accurate proxy for the same.

V. Computationally-efficient algorithms

In this section, we propose a greedy algorithm that approximately solves (15). The accuracy of the algorithm and the new objective function is quantified through numerical experiments in the next section by comparing against the case with full feedback bandwidth.

In general, it is imperative that any feedback optimization be performed while paying only a reasonable price in terms of control overhead for otherwise, the purpose of throughput enhancement through intelligent allocation is defeated. For this reason, we consider the class of *c-sparse* algorithms that allocate full feedback bandwidth to $\lfloor c \rfloor$ carefully-chosen users and the remaining bandwidth to a $\lceil c \rceil$-th chosen user. Note that the control overhead for such a class is exactly $\log_2 \binom{K}{\lceil c \rceil} = \mathcal{O}(\log_2 K)$ bits, when c is a constant, thus scaling gracefully in the number of users. Recall that $c \in (0, K]$ parameterizes the total feedback budget given by $B = cL\log_2 M$. Through the remainder of this paper, we restrict our attention to allocating

$$R_i^F\left(\vec{Q},\vec{\alpha},\vec{g}\right) = \frac{Q_i(\bar{t})}{\sum_{p=1}^{m-1}\chi_p}\sum_m\left[P_{X_{ij}}(r_m)\right]^{2^{g_i}}\left[\sum_{n=m}^{M} r_n\chi_{in}\right]\left[\prod_{k\neq i}\left[1-\left(P_{X_{k1}}\left(\frac{r_m}{\sigma_{ki}}\right)\right)^{2^{g_k}}\right]+\prod_{k\neq i}\left[1-\left(P_{X_{k1}}\left(\frac{r_{m-1}}{\sigma_{ki}}\right)\right)^{2^{g_k}}\right]\right] \quad (15)$$

feedback bandwidths of the form $c \in \{1,2,\ldots,K\}$. In the context of c-sparse algorithms, this has a natural interpretation of allowing the "best" c users to transmit full feedback information back to the base station. Mathematically speaking, the best c-sparse algorithm is one that solves the *c-sparse version* of (4) given by

$$\begin{aligned}\vec{g}(\bar{t})^* = \quad & \arg\max \quad d_F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{g}\right) \\ & \text{s.t.} \quad \sum_{k=1}^{K} 2^{-g_k} \leq c \\ & \quad\; g_k \in \{0,\infty\}.\end{aligned} \quad (16)$$

The naïve brute-force approach to solving (16) incurs complexity $\binom{K}{c}$ operations, which might be too large. For a system with $K = 32$ and $c = 8$ (25% of the full CSI bandwidth), this corresponds to ten million operations. Thus, the question now becomes can we efficiently determine which users deserve the pie, possibly c users that maximize some metric.

To address the user-selection issue, we begin with the simple observation the solution to the c-sparse feedback allocation problem in (16) involves finding the subset of users $\mathcal{S} \subseteq \{1,2,\ldots,K\}$ that maximizes $\sum_{k\in\mathcal{S}} R_k^F\left(\vec{Q}_{\mathcal{S}}(\bar{t}),\vec{\alpha}_{\mathcal{S}}(\bar{t}),c\vec{1}_{\mathcal{S}}\right)$ where $\vec{1}$ is an all-ones vector of length K and $\vec{x}_{\mathcal{S}}$ denotes a restriction of vector $\vec{x}$ to the elements in set $\mathcal{S}$. A useful interpretation of the above process is that we essentially reduce the amount of *competition* by eliminating users $\{1,2,\ldots,K\}\setminus\mathcal{S}$ and then re-evaluate the performance of the c chosen users in this new setting. We then proceed with the intuitive argument that if user k^* ranks best amongst the full field of competition, i.e., $R_{k^*}^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{0}\right) > \max_{j\neq k^*} R_j^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{0}\right)$, then it is unlikely that the same user loses its ranking when the field of competition is reduced. This is because the user wins on the weight of its own merit, in this case, its channels and queue sizes.

The above intuition motivates the following algorithm. The algorithm begins by sorting the users based on $\left\{R_i^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{0}\right)\right\}$, the full feedback queue-weighted drains. We then allocate full feedback bandwidth to the first c users. The algorithm is described more formally below in the general case with $c \in (0,K]$. The proposed formulation

Algorithm 2 Greedy feedback allocation

1: Set $g_i = \infty$ for all i, i.e., initialize to zero feedback bandwidth for all users.
2: Sort $\left\{Q_i(\bar{t})R_i^F\left(\vec{Q}(\bar{t}),\vec{\alpha}(\bar{t}),\vec{0}\right)\right\}$ in descending order. Index the sorted set by $\{i_1,i_2,\ldots,i_K\}$.
3: Set $g_{i_k} = 0$ for $k = 1,2,\ldots,\lfloor c\rfloor$ and $g_{i_{\lceil c\rceil}} = -\log_2(c-\lfloor c\rfloor)$.

in (16) can be seen as a generalization of the formulation proposed by Gopalan et al. [11] to a multi-carrier setting. In particular, when $L = 1$, (16) reduces to the throughput-optimal feedback allocation problem in [11]. The authors do not address the two challenges raised at the end of Section III, that of efficient computation of the objective function in (16) followed by efficient optimization. In contrast, we are able to propose an efficient algorithm above that outputs a c-sparse solution, a development that is made possible by the derivations in the previous section.

In the next section, we run extensive numerically experiments under realistic settings in order to evaluate our proposed two time-scale greedy feedback allocation and data scheduling algorithm.

VI. Numerical experiments

In Section VI-A, we introduce the simulation setup and this is followed by Section VI-B where we present the results of our experiments. We compare the throughput of the greedy feedback allocation algorithm with the full CSI scheduling algorithm as well as against an equal allocation algorithm where each user is given the same amount of feedback bandwidth.

A. Simulation setup

The simulation parameters, that are closely aligned to the LTE specification [20]–[22], are described below:

Network geography: We consider a circular cell of radius $D = 7$km, which is of the order of a typical urban cell-size [22]. The base station is located at the center $\left(\frac{D}{2},\frac{D}{2}\right)$.

Arrivals process: We study both symmetric and asymmetric deterministic arrivals processes in our experiments. The asymmetric case is motivated by the increasing and concurrent demand from mobile users for a plethora of applications such as gaming, video, VoIP, file downloads, etc. We model asymmetric arrivals using parameter κ and set the arrival rate vector to be

$$\vec{\lambda} = [\underbrace{\kappa\lambda\ \kappa\lambda\ldots\kappa\lambda}_{\frac{K}{2}}\ \underbrace{\lambda\ \lambda\ldots\lambda}_{\frac{K}{2}}]^T \text{bits/s},\ \kappa\in\{1,2,5\}. \quad (17)$$

This means that one half of the users in the system are assumed to have larger throughput demands. Of course, $\kappa = 1$ represents the symmetric case. *Without loss of generality, we normalize the duration of a slot to one second and treat (17) as the number of bits arriving in one slot.*

Spatial distribution, path-loss, large-scale coherence time: We assume that the users are uniformly distributed on the circle. The users positions induce a path-loss gain [20], [23] at time t that is given by

$$\alpha_i(t) = 10^{-(32.45+20\log_{10}(f_c)+20\log_{10}(d_i(t)))}.$$

where $d_i(t)$ is the distance from the i-th user to the base station, and $f_c = 1800$MHz denotes the carrier frequency. Recall that users change positions every T_{LS} slots. We set to $T_{LS} = 25$s in our simulations.

Bandwidth, number of OFDMA sub-carriers, transmit power and noise modelling: The total system bandwidth

is $\omega = 10$MHz with $L = 1024$. The transmit power is set to $P = 10$W which is calculated as 46dbm (EIRP) $-$ 6dB (Cable losses) $= 10$dB. The additive noise at the receiver is modelled [20], [23] as

$$N_o = 10^{\frac{-147+NF+10log_{10}(\omega/L)}{10}},$$

where $NF = 9$dB is the noise floor.

Channel distribution and CQI Table: We use a 16-state CQI table from the LTE standard [24]

$$\begin{aligned}\mathcal{M} &= \{0.05, 0.15, 0.23, 0.37, 0.6016, 0.877, 1.17,\\ &\quad 1.47, 1.91, 2.4, 2.7, 3.3, 3.9, 4.5, 5.1, 5.55\}.\end{aligned}$$

Note that as the first CQI value is not specified by the authors [24], we choose an arbitrary value of 0.05bps/Hz. Recall that the channels are assumed to be i.i.d. across OFDMA bands. The probability distribution (on any band)

$$\Pr\left(X_{ij}(t) = r_m\right) = \rho_{mi}\left(\alpha_i\left(\bar{t}\right)\right) \tag{18}$$

is calculated as follows. We discretize/quantize the Shannon capacity random variable given by $\log_2\left(1 + \frac{P\alpha_i(\bar{t})h}{N_o}\right)$, where $h \sim \exp(1)$ is a standard exponential random variable, using codebook $\mathcal{M}$. One can then derive the probability distribution induced on the CQI Table or codebook $\mathcal{M}$ as $\rho_{mi}\left(\alpha_i\left(\bar{t}\right)\right) = \exp\left(-\frac{2^{r_m}-1}{\alpha_i(\bar{t})}\right) - \exp\left(-\frac{2^{r_{m+1}}-1}{\alpha_i(\bar{t})}\right)$ for $m = 1, 2, \ldots, M-1$, and $\rho_{mi}\left(\alpha_i\left(\bar{t}\right)\right) = \exp\left(-\frac{2^{r_m}-1}{\alpha_i(\bar{t})}\right)$ for $m = M$.

Intuitively speaking, one expects sub-carrier grouping to exhibit the worst performance versus the full feedback case when the channel distribution in (18) has lower variance, e.g., in the low-SNR and high-SNR regime. In these regimes, it is likely that $X_{ij}^{eff} \approx X_{ij}$ for any g_i thus alleviating the loss in throughput due to grouping. We are therefore motivated to evaluate the performance of the grouping approach under a uniform spatial distribution on the circle, which translates into a uniform distribution on $\mathcal{M}$ that has highest variance.

Number of users and feedback budget: The number of users is set to $K = 32$. We consider two cases $c \in \{4, 8\}$ for the feedback budget that is given by $B = cL\log_2 M$; $c = 4$ and $c = 8$ corresponds to feedback reductions of 125% and 75% respectively in relation to full feedback bandwidth.

B. Simulation Results

Having described the simulation setup in detail, we now present the results of our experiments. Under each case $(c, \kappa) \in \{4, 8\} \times \{1, 2, 5\}$, we compare the performance of three algorithms: the full feedback case with $c = K$, the greedy feedback algorithm described in Section V and an equal allocation algorithm that is described next.

The equal allocation algorithm is a "one-shot" algorithm that divides the total feedback bandwidth equally amongst all users at the beginning of the communication epoch. In other words, the feedback allocation is given by

$$g_k^{EQ}(t) = g_k^{EQ} = \left\lceil -\log_2\left(\frac{c}{K}\right)\right\rceil, \ \forall k, \forall t. \tag{19}$$

As a sanity check on our simulation results, one can derive a lower bound on the throughput that the above equal allocation algorithm can support. This is made possible by the simple observation that any arbitrary scheduler provides a service rate that can be achieved by MaxWeight owing to the optimality of the latter. The lower bound is calculated based on the max-sum-rate scheduler, which selects the user with the best channel in each OFDMA sub-band. The total system rate in the case of the max-sum-rate scheduler can be calculated in closed-form, using similar arguments as presented earlier in the paper, as

$$\begin{aligned}\textstyle\sum_j \mathbb{E}\left[\max_k X_{kj}^{eff}\right] &= L\mathbb{E}\left[\max_k X_{kj}^{eff}\right]\\ &= L\textstyle\sum_{m=1}^{M} r_m \Pr\left(\max_k X_{kj}^{eff} = r_m\right)\\ &= L\textstyle\sum_{m=1}^{M} r_m \left[\prod_k \left(1 - \Pr\left(X_{k1} > r_m\right)^{2^{g_k}}\right)\right.\\ &\quad \left. - \prod_k \left(1 - \Pr\left(X_{k1} > r_{m-1}\right)^{2^{g_k}}\right)\right],\end{aligned} \tag{20}$$

where $r_0 = 0$. When both the arrival processes and the user channels are symmetric, the supportable per user is at least

$$\begin{aligned}\mu_k^{EQ}(c) &= \tfrac{L}{K}\textstyle\sum_{m=1}^{M} r_m \left[\prod_k \left(1 - \left(1 - \frac{m}{M}\right)^{2^{g_k^{EQ}}}\right)\right.\\ &\quad \left. - \prod_k \left(1 - \left(1 - \frac{m-1}{M}\right)^{2^{g_k^{EQ}}}\right)\right].\end{aligned} \tag{21}$$

In fact, under symmetric arrivals and channels, (21) represents the maximum supportable symmetric throughput under *one-shot* feedback allocation policies with constraint $B = cL\log_2 M$, $c \in \{1, 2, 3, \ldots\}$ bits.

We use the above analysis to perform a sanity check of the results presented in Fig. 1 that correspond to a symmetric setting ($\kappa = 1$) and uniform for $L = 1024$, $K = 32$ and $c \in \{4, 8\}$. In particular, we find that the one-shot equal allocation algorithm with $c = 4$ and $c = 8$ is indeed able to support a maximum throughput of roughly 350kbps and 800kbps respectively as predicted by the above analysis ($\mu_k^{EQ}(6) = 353.56$kbps and $\mu_k^{EQ}(8) = 806$kbps respectively). The dynamic feedback allocation algorithm however outperforms the equal allocation approach by 242% (80% resp.) as shown when $c = 4$ ($c = 8$ resp.). We also see that loss in throughput between the greedy algorithm and the full feedback case is only 13% when $c = 8$. When $c = 4$, the loss in throughput due to sub-carrier grouping is 42%. The results from the asymmetric case ($\kappa = 2$ and $\kappa = 5$) are plotted in Fig. 2. Here, we see that the greedy

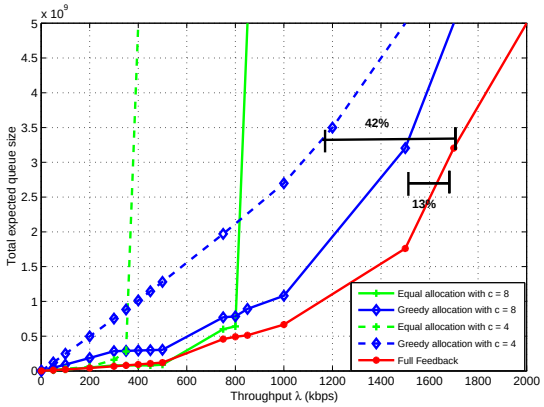

Fig. 1. Throughput under three feedback schemes with symmetric arrivals $\kappa = 1$. The average queue length is measured over 10000 iterations.

algorithm outperforms the equal allocation approach by almost 275% (200% resp.) and underperforms with respect to the full feedback algorithm by 47% (10% resp.) when $c = 4$ ($c = 8$ resp.). Similarly with $\kappa = 5$, we see that the greedy algorithm outperforms the equal allocation approach by 400% (150% resp.) and underperforms with respect to the full feedback algorithm by 30% (10% resp.) when $c = 4$ ($c = 8$ resp.).

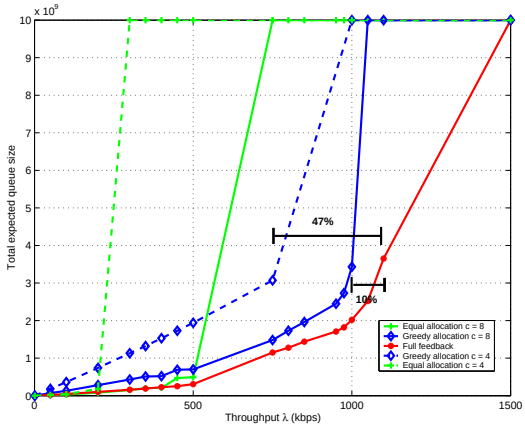

Fig. 2. Throughput under three feedback schemes with asymmetric arrivals $\kappa = 2$. The average queue length is measured over 10000 iterations.

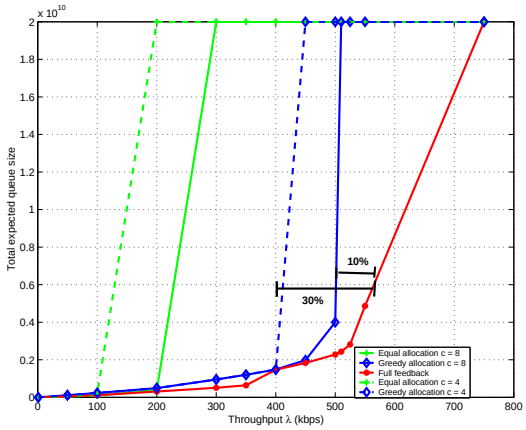

Fig. 3. Throughput under three feedback schemes with asymmetric arrivals $\kappa = 5$. The average queue length is measured over 10000 iterations.

The competitive performance of proposed algorithm leads us to an interesting observation that solutions in the space $\mathcal{G}_{c-sparse} = \{(g_1, \ldots, g_K) : \sum_{k=1}^{K} 2^{-g_k} \leq c,\ g_k \in \{0, \infty\},\ \forall k\}$ are good approximations of the true solutions that lie in $\mathcal{G} = \{(g_1, \ldots, g_K) : \sum_{k=1}^{K} 2^{-g_k} \leq c,\ g_k \in \{0, 1, \ldots, \log_2 L, \infty\},\ \forall k\}$. The gains from greedy feedback allocation over the equal allocation policy do of course come with some expenditure in bandwidth due to control overhead. As mentioned earlier, the required bandwidth scales logarithmically in the number of users and is exactly equal to $\frac{c \log_2 K}{T_{LS}}$ bits per second. When $c \in \{4, 8\}$, $K = 32$, and $T_{LS} = 25$s, the overhead amounts to 0.8bps and 1.6bps, which is negligible.

VII. Concluding Remarks

This paper develops a joint sub-carrier grouping and data scheduling policy that operates on two time-scales. The sub-carrier grouping technique determines a grouping factor for each user that maximizes the queue-weighted drain of the system. We derive a closed-form bound on the queue-weighted drain that is tight for most systems with large packet sizes and sufficient small CQI tables. The closed-form expression in turn allows for the development of a fast greedy algorithm that incurs control overhead that scales only logarithmically in the number of users. We evaluate the performance of the greedy algorithm through numerical experiments. Future directions include developing feedback reduction protocols that involve both sub-carrier grouping and codebook size adaptation.

Appendix A
Proof of Lemma 1

Without loss of generality, we restrict our analysis to the first sub-band in all three parts of the proof since all bands are statistically equivalent.

(a) Without loss of generality, we restrict our analysis to the first sub-band since all bands are statistically equivalent. Firstly, since the channels are independent across users, we can drop the conditioning to get $\Pr\left(X_{i1}^{eff} \geq r_m \,\middle|\, Y_i \in \mathcal{A}_m\right) = \Pr\left(X_{i1}^{eff} \geq r_m\right)$. The remainder of the proof, given below, is immediate since the channels are i.i.d. across sub-bands for a given user

$$\begin{aligned} \Pr\left(X_{i1}^{eff} \geq r_m\right) &= \Pr\left(\min_{s \in \mathcal{F}_i(1)} X_{is} \geq r_m\right) \\ &= \left[P_{X_{ij}}(r_m)\right]^{2^{g_i}}. \end{aligned} \tag{22}$$

(b) The proof proceeds as follows

$$\begin{aligned} &\Pr\left(Y_i \in \mathcal{A}_m\right) \\ &= 1 - \Pr\left(Y_i \notin \mathcal{A}_m\right) \\ &= 1 - \Pr\left(\max_{k \neq i} \sigma_{ki} X_{k1}^{eff}(t) \notin [\, r_{m-1}, r_m\,)\right) \\ &= \Pr\left(\max_{k \neq i} \sigma_{ki} X_{k1}^{eff}(t) < r_m\right) - \Pr\left(\max_{k \neq i} \sigma_{ki} X_{k1}^{eff}(t)\right. \\ &\quad \left. < r_{m-1}\right) \text{since } (0, r_{m-1}) \text{ and } (r_m, \infty) \text{ are disjoint intervals} \\ &= \prod_{k \neq i} \Pr\left(X_{k1}^{eff}(t) < \tfrac{r_m}{\sigma_{ki}}\right) + \prod_{k \neq i} \Pr\left(X_{k1}^{eff}(t) < \tfrac{r_{m-1}}{\sigma_{ki}}\right) \\ &= \prod_{k \neq i} \left[1 - \left(P_{X_{k1}}\left(\tfrac{r_m}{\sigma_{ki}}\right)\right)^{2^{g_k}}\right] \\ &\quad + \prod_{k \neq i} \left[1 - \left(P_{X_{k1}}\left(\tfrac{r_{m-1}}{\sigma_{ki}}\right)\right)^{2^{g_k}}\right], \end{aligned} \tag{23}$$

where the penultimate step follows since the channels are i.i.d. across users on any given frequency band.

(c) The events $\{X_{ij}^{eff} \geq r_m\}$ and $\{Y_i \in \mathcal{A}_m\}$ are independent and thus $\mathbb{E}\left[Q_i(\bar{t}) X_{ij}^{eff}(t) \,\middle|\, X_{ij}^{eff} \geq r_m,\ Y_i \in \mathcal{A}_m\right] = \mathbb{E}\left[Q_i(\bar{t}) X_{ij}^{eff}(t) \,\middle|\, X_{ij}^{eff} \geq r_m\right]$. The latter conditional expectation is straightforward to compute and given as

$$\mathbb{E}\left[Q_i(\bar{t}) X_{i1}^{eff}(t) \,\middle|\, X_{i1}^{eff} \geq r_m\right] = \frac{1}{\sum_{p=1}^{m-1} \chi_p} \sum_{n=m}^{M} r_n \chi_{in}$$

where $\chi_{in} = \Pr\left(X_{i1}^{eff}(t) = r_n\right)$ is the probability distribution of the effective channels of user i. The probability distribution function is derived below

$$\begin{aligned} \chi_{in} &= \Pr\left(X_{i1}^{eff}(t) = r_n\right) \\ &= \Pr\left(X_{i1}^{eff}(t) > r_{n-1}\right) - \Pr\left(X_{i1}^{eff}(t) > r_n\right) \\ &= \begin{cases} \left(P_{X_{i1}}(r_n)\right)^{2^{g_i}} - \left(P_{X_{i1}}(r_{n+1})\right)^{2^{g_i}}, & n = 1, \ldots, M-1 \\ \left(P_{X_{i1}}(r_n)\right)^{2^{g_i}}, & n = M. \end{cases} \end{aligned}$$

The result follows.

REFERENCES

[1] M. Ouyang and L. Ying, "On scheduling in multi-channel wireless downlink networks with limited feedback", *Proc. of the Allerton Conf. on Communication, Control, and Computing*, Monticello, IN, Oct. 2009.

[2] "Overview of 3GPP Release $10V0.0.5(2009-12)$", http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases, Dec. 2009.

[3] 3GPP, "3GPP TS 36.300-870", 3rd Generation Partnership Project.

[4] S. Donthi and N. Mehta, "Joint performance analysis of channel quality indicator feedback schemes and frequency-domain scheduling for LTE", *IEEE Trans. Vehicular Tech.*, vol. PP, pp. 1-13, 2011.

[5] M. -O. Pun, J. K. Kyeong and H. V. Poor, "Opportunistic scheduling and beamforming for MIMO-OFDMA downlink systems with reduced feedback", *IEEE Intern. Conf. on Communications (ICC)*, Bejing, China, May 2008.

[6] D. Gesbert and M.-S. Alouini, "How much feedback is multiuser diversity really worth?", *in Proc. of Intern. Conf. on Communications (ICC)*, pp. 234-238, Jul. 2004.

[7] J. Chen, R. A. Berry, and M. L. Honig, "Limited feedback schemes for downlink OFDMA based on sub-channel groups", *IEEE Journ. Sel. Areas Commun.*, vol. 26, pp. 1451-1461, Oct. 2008.

[8] R. Agarwal, V. Majjigi, Z. Han, R. Vannithamby and J. Cioffi, "Low complexity resource allocation with opportunistic feedback over downlink OFDMA networks", *IEEE Journ. Sel. Areas Commun.*, vol. 26, pp. 1462-1472, Oct. 2008.

[9] W. Dai, B. Rider, and Y. Lui, "Multi-access MIMO systems with finite rate channel state feedback", *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IN, Oct. 2005.

[10] E. Jorswieck, A. Sezgin, B. Ottersten and A. Paulraj, "Feedback reduction in uplink MIMO OFDM systems by chunk optimization", *EURASIP Journal on Advances in Sig. Proc.*, article no. 59, Jan. 2008.

[11] A. Gopalan, C. Caramanis, and S. Shakkottai, "On wireless scheduling with partial channel-state information", *Proc. of the Allerton Conf. on Communication, Control, and Computing*, Monticello, IN, Sep. 2007.

[12] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks", *IEEE Trans. Automatic Control*, vol. 37, pp. 1936-1949, Dec. 1992.

[13] E. Perahia and R. Stacey, "Next Generation Wireless LANs: Throughput, robustness, and reliability in 802.11n", *Cambridge*, Dec. 2009.

[14] H. Ganapathy, S. Banerjee, N. Dimitrov and C. Caramanis, "Optimal feedback allocation algorithms for multi-user uplink", *Proc. of the Allerton Conf. on Communication, Control, and Computing*, Monticello, IN, Oct. 2009.

[15] IEEE 802.16, "Part 16: Air interface for broadband wireless access systems", May 2009.

[16] L. Ying and S. Shakkottai, "On throughput-optimal scheduling with delayed channel state feedback", *In Proc. 2008 Information Theory and Applications Workshop*, pp. 339 - 344, San Diego, CA, Feb. 2008.

[17] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates", *Probability in the Engineering and Informational Sciences*, vol. 18, pp. 191-217, Apr. 2004.

[18] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation", *INFORMS*, vol. 53, pp. 12-25, Jan. 2005.

[19] M . Andrews and L. Zhang, "Scheduling algorithms for multicarrier wireless data systems", *IEEE/ACM Trans. Networking*, vol. 19, pp. 447 - 455, Apr. 2011.

[20] H. Holma and A. Toskala, "LTE for UMTS, OFDMA and SC-FDMA based radio access", *John Wiley & Sons*, 2009.

[21] H. Holma and A. Toskala, "WCDMA for UMTS: HSPA Evolution and LTE", *John Wiley & Sons*, 2010.

[22] E. Dahlman. S. Parkvall, J. Skold and P. Beming, "3G Evolution: HSPA and LTE for Mobile Broadband", *Academic Press*, 2nd edition, 2008.

[23] "Techniques and trends in signal monitoring, frequency management and geolocation of wireless emitters", *Aglient whitepaper*, www.home.agilent.com/agilent.

[24] M. Rumney, "LTE and the evolution to 4G wireless: Design and measurement challenges", *Agilent Technologies*, June 2009.