Self-organizing Dynamic Fractional Frequency Reuse for Best-Effort Traffic Through Distributed Inter-cell Coordination

Alexander L. Stolyar
Bell Labs, Alcatel-Lucent
Murray Hill, NJ 07974
stolyar@research.bell-labs.com

Harish Viswanathan
Bell Labs, Alcatel-Lucent
Murray Hill, NJ 07974
harishv@research.bell-labs.com

***Abstract*—Self-optimization of the network, for the purposes of improving overall capacity and/or cell edge data rates, is an important objective for next generation cellular systems. We propose algorithms that automatically create efficient, soft fractional frequency reuse (FFR) patterns for enhancing performance of orthogonal frequency division multiple access (OFDMA) based cellular systems for forward link best effort traffic. The Multi-sector Gradient (MGR) algorithm adjusts the transmit powers of the different sub-bands by systematically pursuing maximization of the overall network utility. We show that the maximization can be done by sectors operating in a semi-autonomous way, with only some gradient information exchanged periodically by neighboring sectors. The Sector Autonomous (SA) algorithm adjusts its transmit powers in each sub-band independently in each sector using a non-trivial heuristic to achieve out-of-cell interference mitigation. This algorithm is completely autonomous and requires no exchange of information between sectors. Through extensive simulations, we demonstrate that both algorithms provide substantial performance improvements. In particular, they can improve the cell edge data throughputs significantly, by up to 66% in some cases for the MGR, while maintaining the overall sector throughput at the same level as that achieved by the traditional approach. The simulations also show that both algorithms lead the system to "self-organize" into efficient, soft FFR patterns with no a priori frequency planning.**

I. Introduction

Fourth generation cellular systems are currently being developed and will be deployed in a few years time. These systems target significantly higher sector capacities and higher per user data rates compared to third generation systems. In particular, one of the goals of these systems is to boost performance of users at the cell edge that typically suffer from significant out-of-cell interference. Having approached the information-theoretic limits of point-to-point communication through coding and multiple input multiple output (MIMO) techniques, further advances in cellular performance require focusing the attention on efficiently eliminating interference.

In [11] we proposed a self-organizing interference avoidance scheme for constant bit rate traffic in orthogonal frequency division multiple access (OFDMA) systems through selfish optimization of resources by each sector, and demonstrated that efficient fractional frequency reuse (FFR) patterns could be achieved dynamically. In a similar vein, in this paper, we propose algorithms for improving the throughput performance for best effort traffic in OFDMA cellular systems through formation of FFR patterns automatically. We propose two different algorithms, namely the *Multi-sector Gradient* (MGR) that requires some information to be exchanged between neighboring sectors, and *Sector Autonomous* (SA) that is completely distributed and requires no exchange of information.

MGR algorithm adjusts the transmit powers of the different sub-bands by systematically pursuing local maximization of the overall network utility. We show that the maximization can be done semi-autonomously by each sector with only periodic exchange between interfering sectors of a few key variables that naturally arise from the optimization approach. The computations are still distributed and performed independently in each sector.

SA algorithm, on the other hand, adjusts transmit powers in each sub-band independently in each sector using a non-trivial heuristic to achieve out-of-cell interference mitigation. This algorithm is completely autonomous and requires no exchange of information between sectors. Such an algorithm may be desirable when it is not possible to exchange any information between the relevant sectors. MGR, of course, outperforms the SA algorithm.

Both MGR and SA algorithms are only concerned with the power allocation (and reallocation) among the sub-bands by each sector, which is done on a relatively slow time scale. Given the power levels set by either algorithm, each sector can perform an opportunistic, channel-aware scheduling, taking advantage of the fast fading by proper assignment of users to sub-bands (on the fast time scale). We demonstrate through simulations that the performance of MGR and SA algorithms, when compared to that of the standard "universal reuse" (UNIVERSAL) approach where equal powers are assigned to each sub-band in each sector and channel-aware fast time scale scheduling is utilized within each sector, is significantly better especially in increasing cell edge user throughputs. This is despite the fact that when the channel fading is present, *any* power allocation approach, even equal power allocation across the sub-bands as in the UNIVERSAL algorithm, benefits from some level of interference avoidance due to fast channel-aware

scheduling and proper fast (re)assignment of users to sub-bands. (See [12].) The other main contribution of this paper is that, as part of MGR approach, we propose and rigorously substantiate a novel – "virtual scheduling" – algorithm, which allows efficient real-time computation by each sector of the gradient of the system utility function with respect to the current sub-band transmit powers in the sector.

The paper is organized as follows. In Section I-A we briefly discuss some related work. In Section II we describe the system model under consideration and provide the overview of the proposed algorithms. Section III defines the MGR algorithm, with Sections IV and V addressing its key part - the virtual scheduling algorithm for the utility gradient estimation. In Section VI we define the SA algorithm. Numerous simulation studies comparing the performance of MGR, SA and UNIVERSAL algorithms in a realistic setting are given in Section VII. In Section VIII, we then illustrate through simulations the near global optimality of the MGR algorithm performance. We conclude with a discussion of future work in Section IX.

A. Related Work

Numerous papers have been published on scheduling in OFDMA systems. However, most of these papers are focused on single cell scheduling and typically do not consider the effect of out-of-cell interference. Several papers [3], [5], [7] have been published on coordinated scheduling, although not in the context of OFDMA. These papers propose algorithms that are centralized and are not based on simple exchange of messages between sectors as in this paper. Dynamic distributed resource allocation in the context of Gaussian interference channels has been considered in [1] and [4]. Neither of these papers considers the model of this paper with multiple interfering base stations each serving several, differently located users. (As a result, in our model, even within the same sector, different users experience different interference levels in different sub-bands.) The concept of FFR for best effort traffic in the context of OFDMA systems has appeared in cellular network standardization technical contributions [13], [14] and in [6]; a scheme conceptually close to FFR was proposed in [8], but for a model different from ours and oriented towards a larger time-scale (hours) optimization whereas our goal is to find a scheme that is adaptive on a smaller time-scale (seconds). As mentioned earlier, we proposed and studied a self-organizing FFR scheme for constant bit rate traffic such as voice over Internet Protocol (VoIP) in our prior work [11].

II. System model

A. OFDMA description and key assumptions

We begin with a very brief description of an OFDMA system from [11]. In an OFDMA system the transmission band is divided into a number of sub-carriers and information is transmitted by modulating each of the sub-carriers. Further, time is divided into slots consisting of a number of OFDM symbols and transmissions are scheduled to users by assigning a set of sub-carriers on specific slots. The frequency resources scheduled are usually logical sub-carriers. The logical sub-carriers are mapped to physical sub-carriers via a frequency hopping, which is employed to achieve interference averaging.

OFDMA systems supporting FFR for interference mitigation divide frequency resources into several *sub-bands*. Frequency hopping of sub-carriers is restricted to be within the sub-band so that users scheduled on a certain sub-band experience interference only from transmissions in neighboring sectors in the same sub-band. "Soft" fractional frequency reuse can be achieved by reusing same frequency sub-bands in neighboring sectors, but at different power levels, in a manner that reduces inter-sector interference. Note that sub-band is a special case of a *resource set* which could be a combination of a set of sub-carriers in frequency and a set of time-slots. FFR schemes, including those in this paper, can be implemented using resource sets instead of sub-bands.

Another important aspect of the system, which is assumed in the model described below, is the channel quality indicator feedback from the mobiles. The feedback is used by the channel aware scheduler to select users for each of the sub-bands for transmission in each slot, and also to determine the modulation and coding format. For this purpose, relatively frequent channel quality feedback is required. In addition, relatively infrequent, average signal-to-interference-and-noise ratio (SINR) feedback for each sub-band is also required for our algorithms. Another infrequent feedback, that is unique to MGR algorithm, is the pathloss ratio between the signal and interfering base stations.

B. Formal model

We have K cells (sectors) $k \in \mathcal{K} = \{1, \dots, K\}$, and J sub-bands $j \in \mathcal{J} = \{1, \dots, J\}$ in the system. We assume that each sub-band consists of a fixed number c of sub-carriers, and denote by W the bandwidth of one sub-band. The noise spectral density is denoted by N_0.

Time is slotted, so that transmissions within each cell are synchronized, and do not interfere with each other. A transmission in a cell, assigned to a sub-band, causes interference to only those users in other cells, that are assigned to the same sub-band; there is no inter-sub-band interference.

The *utility* $\mathcal{U}$ of the system (or network) is defined as the sum

$$\mathcal{U} = \sum_k U^{(k)}$$

of utilities $U^{(k)}$ of all sectors. In turn, sector k utility $U^{(k)}$ is a smooth concave function of the *average rates* X_i of users i served by the sector k. The precise conditions on a sector utility function will be specified in Section IV; for example, it can be $U^{(k)} = \sum_i \log X_i$ (with the summation over users i within sector k).

We denote by $P_j^{(k)}$ the power allocated in sub-band j of sector k. The total power within each sector is upper bounded by P^*, so that $\sum_j P_j^{(k)} \le P^*$.

The system objective is to maximize the total utility $\mathcal{U}$, by setting and adjusting the power levels $P_j^{(k)}$. The exact

solution to this problem is very difficult to obtain, even using centralized schemes, as the problem is "highly non-convex." In addition, any practical algorithm should involve very limited real-time information exchange (signaling) among sectors.

The two different algorithms, MGR and SA, that we propose in this paper are such that the power levels $P_j^{(k)}$ are adjusted over time (relatively slowly) with the purpose of improving the system utility, given current set of the users in the system and their current sector assignments. MGR tries to imitate the gradient ascend method, and involves some inter-sector/cell information exchange. Our main contribution in MGR is the *virtual scheduling* algorithm, which constantly estimates the partial derivatives $\partial\mathcal{U}/\partial P_j^{(k)}$ in a very efficient and "distributed" way. Algorithm SA does *not* involve any inter-sector/cell signaling, and is based on reasonable (but not straightforward) heuristics.

III. MGR: Dynamic power allocation algorithm with base station coordination

We now describe the MGR algorithm, according to which sectors dynamically allocate/reallocate the power levels among sub-bands. The algorithm involves sectors (base stations) exchanging information on how "costly" to their utility is the interference caused by other sectors. (We describe the algorithm as if each sector shares this information with *all* other sectors; in reality, and in our simulations, each sector exchanges information only with a small number of its neighboring sectors.)

The idea of the algorithm is simple. Each sector k constantly adjusts its power allocation to different sub-bands in a way that improves the total utility $\mathcal{U} = \sum_m U^{(m)}$ of the system.

MGR ALGORITHM (SUB-BAND POWER ADJUSTMENT PART):

Each sector $k \in \mathcal{K}$ maintains the *estimate* of the utility $U^{(k)}$ which the sector *could potentially* attain, given its current power allocation among sub-bands, $P_j^{(k)}$, $j \in \mathcal{J}$, $\sum_j P_j^{(k)} \leq P^*$, and current interference level from other sectors. Moreover, sector k maintains estimates of partial derivatives $D_j^{(m,k)} = \partial U^{(k)}/\partial P_j^{(m)}$ of its (maximum attainable) utility on the power levels $P_j^{(m)}$ in all sectors m (including self, $m = k$) and all sub-bands j. The key part of the algorithm, and our key contribution, is *how* these estimates are computed; the *virtual scheduling* algorithm which does that is described in detail in Section V (which in turn relies on the results of Section IV).

Sector k periodically sends values of $D_j^{(m,k)}$, for all j, to each sector $m \neq k$. Correspondingly, it also periodically *receives* the values of $D_j^{(k,m)}$, for all j, from each sector $m \neq k$. (The frequency of such exchange does *not* have to be high.)

Sector k maintains the current values of

$$D_j^k = \sum_m D_j^{(k,m)}, \quad \text{for each sub-band } j. \tag{1}$$

Clearly, D_j^k is the estimate of the partial derivative $\partial\mathcal{U}/\partial P_j^{(m)}$.

In each physical time slot (or more generally, every n_p physical slots), sector k does the following. We use fixed parameter $\Delta > 0$, and denote by $P^{(k)} = \sum_j P_j^{(k)}$ the current total power in the sector. Then, the powers updated, sequentially, as follows:

1. We pick j_* (if such exists) such that $D_{j_*}^k$ is the smallest among those j with $D_j^k < 0$ and $P_j^{(k)} > 0$, and do

$$P_{j_*}^{(k)} \doteq \max\{P_{j_*}^{(k)} - \Delta, 0\}.$$

2. If $P^{(k)} < P^*$, we pick j^* (if such exists) such that $D_{j^*}^k$ is the largest among those j with $D_j^k > 0$, and do

$$P_{j^*}^{(k)} \doteq P_{j^*}^{(k)} + \min\{\Delta, P^* - P^{(k)}\}.$$

3. If $P^{(k)} = P^*$ and $\max_j D_j^k > 0$, we pick a pair (j_*, j^*) (if such exists) such that $D_{j^*}^k$ is the largest, $D_{j_*}^k$ is the smallest among those with $P_j^{(k)} > 0$, and $D_{j_*}^k < D_{j^*}^k$. Then,

$$P_{j_*}^{(k)} \doteq \max\{P_{j_*}^{(k)} - \Delta, 0\},$$

$$P_{j^*}^{(k)} \doteq P_{j^*}^{(k)} + \min\{\Delta, P_{j_*}^{(k)}\}.$$

The initial values are $P_j^{(k)} = P^*/J$. The algorithm runs "continuously", and, therefore, the choice of initial state - at the system start-up or reset - is not crucial.

END ALGORITHM

We want to emphasize the fact that the power adjustment algorithm, as well the virtual scheduling algorithm (being its part), works with estimated maximal possible utility a sector can potentially attain (given current power levels), and not the actual current utility. If power allocations in the system converge, and stay approximately constant, then the "virtual utilities," used by the algorithms run in sectors, will be close to actual ones. However, a real system is dynamic, with users arriving, departing, and moving from sector to sector. As a result, the actual sector utilities can "lag behind" the optimal ones for the current power levels. Virtual utilities estimate the optimal utilities, and thus better determine the desired directions of power adjustments.

IV. Differentiability of a sector utility function on available transmission rates

In this section we consider a fixed sector k, and study the dependence of its utility U on the rates R_{ij}, where R_{ij} is the rate available to user i (*in this sector*) in sub-band j, if this user is chosen for transmission in a time slot. (We assume that rates R_{ij} do not change with time.) More specifically, we derive the expression for the partial derivative $(\partial/\partial R_{ij})U$. To simplify the notation, within this Section IV, we suppress sector index k in the variables, including $U^{(k)}$.

The users in the sector are indexed by $i \in \mathcal{I} = \{1, \ldots, N\}$. In each time slot, for each sub-band one user is chosen to transmit data to; $R_{ij} \in [0, B]$, $B < \infty$, is the transmission rate in sub-band j to user i, if this user is chosen. We will denote $R = \{R_{ij},\ i \in \mathcal{I},\ j \in \mathcal{J}\}$. A scheduling algorithm runs over many time slots. Denote by $\phi_{ij} \in [0, 1]$ the fraction of

time an algorithm chooses user i for transmission in sub-band j. (A scheduling algorithm does not have to - and typically does not - allocate those fractions explicitly; typically, they are what they turn out to be under the algorithm.) Naturally, $\sum_i \phi_{ij} \leq 1, \ \ \forall j$. Then the average rate user i actually receives is

$$X_i = \sum_j \phi_{ij} R_{ij}, \quad \forall i. \tag{2}$$

Given R, the set of all vectors $X = (X_1, \ldots, X_N)$ for all possible $\phi = \{\phi_{ij}, \ i \in \mathcal{I}, \ j \in \mathcal{J}\}$, is a convex compact set $V = V(R)$ in the positive orthant $\mathbb{R}_+^N$. Clearly, $X_i \in [0, JB]$ for all i, for any $X \in V(R)$ and any R.

In this paper, let us assume that the utility function $U(X)$ of the average rate vector X is $U(X) = \sum_i \log X_i$. (It can be a far more general concave function – see [12].)

For each R and corresponding region $V(R)$, consider the unique vector

$$X(R) = \arg\max_{X \in V(R)} U(X).$$

The uniqueness follows from convexity of $V(R)$ and strict concavity of U.

The question is: what is the expression for $(\partial/\partial R_{ij})U(X(R))$? To gain intuition, consider a ϕ corresponding to $X(R)$, i.e. ϕ satisfying (2) with $X(R)$ in place of X. Then, if we formally differentiate $U(X(R))$ on R_{ij}, using (2) and assuming ϕ is constant, we obtain

$$\frac{\partial}{\partial R_{ij}} U(X(R)) = \frac{\partial U}{\partial X_i}(X(R))\phi_{ij}. \tag{3}$$

This is not a proof, of course, and in fact (3) does not always hold. However, we can prove that "typically", (3) does hold. The formal result (proved in [12] as Theorem 10.4) is as follows.

Theorem 4.1: For almost all (with respect to Lebesgue measure) R in $[0, B]^{NJ}$, $U(X(R))$ is continuously differentiable in a neighborhood of R, with partial derivatives given by (3).

V. Sensitivity of a sector utility to power changes

A. General expressions

As in Section IV, we consider a fixed sector k, and use the same notations with suppressed index k: the users $i \in \mathcal{I} = \{1, \ldots, N\}$ are those in the sector; their average throughputs are X_i; R_{ij} are per-user, per-sub-band rates (nominal, i.e., if user is selected); $U(X)$ is the sector utility function, defined also as in Section IV. However, for the per-sector, per-sub-band powers $\{P_j^{(m)}, \ j \in \mathcal{J}, m \in \mathcal{K}\}$ we will retain the sector index m.

Let us denote by $G_i^{(m)}$ the propagation gain from sector m to user i. For the purposes of determining the sensitivity of the sector utility to power changes we assume that the propagation gains are *not* dependent on the sub-band. The values $G_i^{(m)}$ represent the channel gains averaged over the fast fading. This is because the goal of the algorithm is to adapt the transmit power levels to average interference levels and not to track the fast fading. Correspondingly, the instantaneous rates R_{ij} are the rates *as they would be* if the channel gains $G_i^{(m)}$ were constant.

Our goal is to derive expressions for the partial derivatives $\partial/\partial P_j^{(m)}[U(X)]$ for a sub-band j and all sectors $m \in \mathcal{K}$, including $m = k$. We have the following general expression (using (3) and assuming the set R is "typical" in the sense of Theorem 4.1):

$$D_j^{(m,k)} \doteq \frac{\partial}{\partial P_j^{(m)}} U(X) = \sum_i \frac{\partial U}{\partial X_i}(X)\phi_{ij}\frac{\partial R_{ij}}{\partial P_j^{(m)}}. \tag{4}$$

Thus, we need expressions for $\partial R_{ij}/\partial P_j^{(m)}$. We use Shannon formula for the rate

$$R_{ij} = H(F_{ij}(P)), \tag{5}$$

where N_0 is noise spectral density and W is the sub-band bandwidth, and

$$H(y) \doteq W \log_2(1+y), F_{ij}(P) \doteq \frac{G_i^{(k)} P_j^{(k)}}{N_0 W + \sum_{m \neq k} G_i^{(m)} P_j^{(m)}}$$

and $G_i^{(m)}$ is the propagation gain from sector m to user i. Thus,

$$\frac{\partial R_{ij}}{\partial P_j^{(m)}} = H'(F_{ij}(P))\frac{\partial F_{ij}(P)}{\partial P_j^{(m)}}. \tag{6}$$

Finally, given the form of function F_{ij}, we easily obtain

$$\frac{\partial F_{ij}(P)}{\partial P_j^{(k)}} = \frac{F_{ij}(P)}{P_j^{(k)}}, \tag{7}$$

$$\frac{\partial F_{ij}(P)}{\partial P_j^{(m)}} = -\frac{[F_{ij}(P)]^2}{P_j^{(k)}}\frac{G_i^{(m)}}{G_i^{(k)}}, \quad \text{if } m \neq k. \tag{8}$$

The important observation about (6)-(8) is that these expressions *can be easily evaluated by the sector k controller*, because the values of $P_j^{(k)}$ and $F_{ij}(P)$ are directly available to it, and the ratios $\frac{G_i^{(m)}}{G_i^{(k)}}$ of propagation gains for each user i can be evaluated by the user (from the pilot power measurements) and reported to the controller.

B. Virtual scheduling to estimate sensitivity to power changes

In Section V-A we have shown that the sensitivity of a sector k utility to changes in power levels $P_j^{(m)}$ (in all sectors m and sub-bands j), is "typically" given by (4), where the values of partial derivatives $\partial R_{ij}/\partial P_j^{(m)}$ in the RHS are available to sector k controller. The question remains, how the controller can compute or estimate the optimal values of the fractions ϕ_{ij}, maximizing the sector utility $U(X)$? These fractions are hard to find analytically.

Our approach is as follows. To estimate and update the values of partial derivatives $D_j^{(m,k)}$ in (4), for all m and j "simultaneously," each sector k continuously runs a *virtual scheduling* algorithm which is known to (asymptotically)

maximize the sector utility. This is a well-known *gradient scheduling algorithm* (see [9] and references therein). In the special case of $U(X) = \sum_i \log X_i$, it is the *proportional fair* algorithm.

MGR ALGORITHM (VIRTUAL SCHEDULING AND $D_j^{(m,k)}$ ESTIMATION):

The algorithm is run by each sector k independently, over a sequence of "virtual time slots." (The algorithm runs a fixed number n_v of virtual slots within each physical time slot of the system. The greater the n_v the greater the accuracy of the algorithm and its responsiveness to changes in system state; but, the computational burden is greater as well.) The algorithm maintains the current values X_i of average user (virtual) throughputs, and current values of $D_j^{(m,k)}$. It uses small averaging parameters $\beta_1, \beta_2 > 0$, which are chosen in conjunction with n_v. As a general rule, as n_v changes, the product $\beta_j n_v$ has to be kept constant.

In each virtual time slot, we sequentially pick each sub-band j and perform the following steps.

1. Choose user i^*,

$$i^* \in \arg\max_i \frac{\partial U}{\partial X_i}(X) R_{ij}.$$

2. Update:

$$X_{i^*} = \beta_1 J R_{i^* j} + (1-\beta_1) X_{i^*},$$

$$X_i = (1-\beta_1) X_i, \quad \text{for all } i \neq i^*.$$

3. For each m (including $m = k$, that is the sector itself), we update:

$$D_j^{(m,k)} = \beta_2 \frac{\partial U}{\partial X_{i^*}}(X) \frac{\partial R_{i^*,j}}{\partial P_j^{(m)}} + (1-\beta_2) D_j^{(m,k)}. \quad (9)$$

The initial values of the variables are chosen in some arbitrary, but reasonable way (so that their absolute values are not much larger than "correct" values). For example, $X_i = (1/N)\sum_j R_{ij}$ and all $D_j^{(m,k)} = 0$. The algorithm runs "continuously", and, therefore, the choice of initial state - at the system start-up or reset - is not crucial.)

END ALGORITHM

Remark. In the case when the actual scheduling algorithm (described in Section VII-A2) has non-zero minimum rate requirement, the terms $\frac{\partial U}{\partial X_i}(X)$ in the above virtual scheduling algorithm are everywhere replaced by $\exp(aT_i)\frac{\partial U}{\partial X_i}(X)$, where the factor $\exp(aT_i)$ is fed from the actual scheduler.

VI. Sector autonomous power allocation algorithm

A. A discussion of why a version of MGR, but without coordination, does not work

Suppose that for some reason (standards constraints, performance constraints, etc.) inter-cell coordination which is a part of MGR is impossible or undesirable. Then, a natural question is: What if we run a version ("special case") of MGR, but exclude inter-cell coordination? Namely, suppose each sector k estimates only the values of $D_j^{(k,k)}$ (see (9)), that is, sensitivities of its utility to its "own" powers $P_j^{(k)}$; and it uses $D_j^k = D_j^{(k,k)}$ instead of (1). One might hope that such an algorithm, let us call it Single-cell Gradient (SGR), will still result in substantial performance improvement over UNIVERSAL (even if its performance is worse than that of MGR). Unfortunately, this is not the case: SGR typically does not produce a good fractional frequency reuse pattern, and instead has the tendency to equalize powers across sub-bands in most sectors; thus, it typically reverts to UNIVERSAL. This phenomenon is explained, using a simple illustrative example, in [12]. (It's omitted here to save space.)

B. SA: A different algorithm without coordination

Still, the idea of having a completely distributed (with no inter-sector communication) algorithm, producing good FFR patterns and outperforming UNIVERSAL, is very attractive. We will now propose such an algorithm, and call it Sector Autonomous (SA). Although this algorithm does not explicitly maximize the sector utility itself, we believe that it is based on reasonable heuristics. We will show by simulations that its performance is good, (although, as expected, not as good as that of MGR); this algorithm may be an attractive option for applications where extra inter-cell communication is undesirable or infeasible.

The idea of SA is this. We will make each sector to selfishly solve a somewhat different, "artificial" optimization problem, which is however, (a) "highly correlated" with the original one and (b) inherently "encourages" an uneven power allocation to sub-bands (when such is beneficial).

Namely, let us "pretend" that a sector operates in the following way. (We are talking about a single sector, and will suppress sector index k.) Suppose a parameter $\bar{P}$, $P^*/J \leq \bar{P} \leq P^*$, is fixed. In each (virtual) time slot, in each sub-band j, sector either serves (transmits to) exactly one of the users i at power level $\bar{P}$ (and then the transmission rate is R_{ij}, depending on the *actually measured* SNR of user i), or does not serve any user at all (in which case the power used is 0). Now, given this setting, suppose that we employ a scheduling strategy which, over time, solves the following problem: Maximize $\sum_i U_i(X_i)$, where X_i are users' average throughputs, subject to the constraint on the total average power

$$\sum_i \bar{P}_j \leq P^*,$$

where $\bar{P}_j$ is the *average* power (per virtual slot) allocated in sub-band j. This problem is efficiently solved by a virtual scheduling algorithm described below, which runs continuously. (The algorithm is a special case of Greedy Primal-Dual algorithm [10].) Then, the *actual* per-sub-band power levels P_j are set and adjusted to be equal to the average powers $\bar{P}_j$ (continuously produced and adjusted by the virtual scheduling).

SA ALGORITHM: VIRTUAL SCHEDULING FOR $\bar{P}_j$ CALCULATION:

The algorithm is run by each sector k independently, over a sequence of "virtual time slots." (The algorithm runs a fixed

number $n_v \geq 1$ of virtual slots within each physical time slot of the system. The greater the n_v the greater the accuracy of the algorithm and its responsiveness of to changes in system state; but, the computational burden is greater as well.) The algorithm maintains the current values X_i of average user (virtual) throughputs, the current values of $\bar{P}_j$, and a variable Z. It uses a small (averaging) parameter $\beta > 0$, which is chosen in conjunction with n_v. (As a general rule, as n_v changes, the product βn_v has to be kept constant.)

In each virtual time slot, we sequentially pick each sub-band j and do the following.

IF $\max_i \frac{\partial U}{\partial X_i}(X) J R_{ij} - \beta Z \bar{P} \geq 0$,

1a. Choose user i^*,

$$i^* \in \arg\max_i \frac{\partial U}{\partial X_i}(X) R_{ij}.$$

2a. Update:

$$X_{i^*} = \beta J R_{i^* j} + (1-\beta) X_{i^*},$$

$$X_i = (1-\beta) X_i, \quad \text{for all } i \neq i^*,$$

$$\bar{P}_j = \beta \bar{P} + (1-\beta)\bar{P}_j,$$

$$Z = Z + \bar{P}.$$

ELSE

2b. Update:

$$X_i = (1-\beta) X_i, \quad \text{for all } i,$$

$$\bar{P}_j = (1-\beta)\bar{P}_j.$$

END

3. Update:

$$Z = \max\{Z - P^*/J, 0\}.$$

The initial values of the variables are, for example, as follows: $X_i = (1/N)\sum_j R_{ij}$, $\bar{P}_j = P^*/J$, $Z = 0$. (The algorithm runs "continuously", and, therefore, the choice of initial values - at the system start-up or reset - is not crucial.)

END ALGORITHM

Remark. In the case when the actual scheduling algorithm (described in Section VII-A2) has non-zero minimum rate requirement, the terms $\frac{\partial U}{\partial X_i}(X)$ in the above virtual scheduling algorithm are everywhere replaced by $\exp(aT_i)\frac{\partial U}{\partial X_i}(X)$, where the factor $\exp(aT_i)$ is fed from the actual scheduler.

VII. Simulations

A. System model for simulations.

We consider a hexagonal grid of 19 base stations each with three sectors. The sector antennas are assumed to be oriented in a clover-leaf pattern so that the adjacent cell sectors are not facing each other directly. A wrap-around model for interference where the hexagonal arrangement is replicated by translation to create the same number of interfering cells around every one of the 19 cells is adopted. The propagation parameters used are quite standard – they are listed in Table I.

Parameter	Assumption
Cell Layout	Hexagonal 57 sector
Inter-site distance	2.5 Km
Path Loss Model	$L = 133.6 + 35\log_{10}(d)$
Shadowing	Log Normal with 8.9 dB Std. Dev.
Penetration Loss	10 dB
Noise Bandwidth	1.25 Mhz
BS Power	40 dBm
BS Antenna Gain	15 dB
Rx Antenna Gain	0 dB
Rx Noise Figure	7 dB
Channel Model	No fading, Frequency-selective fading

TABLE I
PROPAGATION PARAMETER VALUES USED IN THE SIMULATIONS

We simulate an OFDMA system with 48 sub-carriers divided into 6 sub-bands with the same number of sub-carriers in each sub-band. The time-slot is 1 msec; all simulations are run over 5000 slots. The system load is 20 users per sector. The full buffer traffic model is used for all the simulation results in this paper, i.e. all users have an infinite amount of back-logged traffic. (In [12] we also present results for a bursty traffic.)

To demonstrate the effect of fast fading we run the simulations with and without it. The model of fast fading is representative of frequency-selective Rayleigh fading with temporal characteristics captured through Jakes model with vehicle speed of 20 Km/hr and carrier frequency of 2 Ghz. The frequency-selectivity is modeled by simulating independent fading across three sets of coherence bands each comprising two sub-bands (block frequency fading model).

1) Transmit power allocation: The transmit power of each sub-band is determined by the algorithm in Section III for the MGR algorithm and by the algorithm in Section VI for the SA algorithm.

In the case of MGR, the virtual scheduling algorithm described in Section V-B is run every slot. The number of virtual slots is set at 30. The various parameter values used in the virtual scheduling algorithm are $\beta_1 = 0.005, \beta_2 = 0.01$. The values of the rates R_{ij} and of the gain ratios $\frac{G_i^{(m)}}{G_i^{(k)}}$ used by the virtual scheduling are computed based on the signal-to-interference-and-noise ratio (SINR) feedback from mobiles; these values are averages over roughly 500 slots, and thus change slowly from slot to slot. (The details of the calculations are given in [12], and omitted here to save space.)

In the case of SA the virtual scheduling algorithm described in Section VI is also run every slot with 30 virtual slots. The various parameter values are $\beta = 0.01, J = 6, \bar{P} = (2/3)P^*$.

2) Actual scheduling of transmissions: With the powers $P_j^{(k)}$ for each sub-band in sector k dynamically determined by the appropriate algorithm (either MGR or SA), actual scheduling is implemented independently by each sector k. The scheduling algorithm is such that it maximizes the utility $U^{(k)}$ of the sector, *given the current power-to-sub-band allocation in the system.* (This obviously means that the total

system utility under the current power allocation is maximized as well.) We use the utility function $U^{(k)} = \sum_i \log(\bar{X}_i)$, where the summation is over users i served by sector k, and $\bar{X}_i$ are users' *actual* average throughputs. (These are generally *not* the average throughputs X_i used in the virtual scheduling algorithms.) This utility function results in the well known *proportional fair* scheduling (cf. [2]). In each time slot, the potential *instantaneous* rates, $\hat{R}_{ij}$, are determined (based on SINR feedback) by the serving sector for all its users in all sub-bands. Then, in this slot, in each sub-band j a user with the maximum value of the metric $\hat{R}_{ij}/\bar{X}_i$ is scheduled (and assigned the entire sub-band). The average rates $\bar{X}_i$ are updated only upon successful transmission of the packets of corresponding users.

In our simulations we, in fact, use a generalization of the proportional fair scheduling algorithm (see [2]) which allows us to introduce minimum rate requirements of the form $\bar{X}_i \geq b$ for some constant $b \geq 0$. The generalized algorithm maintains a *token counter* variable T_i for each user i, and uses a more general scheduling metric of the form $\exp(aT_i)\hat{R}_{ij}/\bar{X}_i$, where $a > 0$ is a parameter. The factor $\exp(aT_i)$, maintained by the actual scheduler, is also fed to and used by the virtual scheduling algorithms (see remarks in Sections V-B and VI-B.)

B. Results and discussion

To illustrate that both the MGR and SA algorithms create soft fractional frequency reuse patterns automatically, we show in Figures 1 and 2 the slot by slot dynamics of transmit power allocation in each of the six sub-bands. Initially (in slot 0), equal power is allocated to all sub-bands in all sectors. The powers are shown for three out of the 57 sectors that are roughly facing each other. The results are for the case of the frequency-selective fading channel and uniform user distribution. The figures clearly show that both algorithms dynamically adjust powers and allocate them unequally among sub-bands. It is also clear that the reuse pattern achieved is a soft reuse in the sense that all sub-bands are used in all sectors but with different power levels. Such a reuse pattern, in turn, depends on the system layout, user distribution, propagation gains, etc. We remark that, although it may appear that power levels "never quite converge," we have observed that such "jitter" in power allocations has very little effect on the achieved value of system utility (as will be defined shortly), which remains stable after the initial transience period.

Simulation results comparing the performance of the 3 different algorithms, namely MGR, SA, and UNIVERSAL, are presented in the form of geometric average of user throughputs versus the 5-percentile throughput. We use the geometric average throughput (GAT) as the performance metric, because maximizing it is the algorithms' objective (recall that the utility function is the sum of log-throughputs), and it is easier to "relate to" than the sum of log-throughputs metric. In particular, percentage improvements are much more meaningful in the GAT metric than the sum of log-throughputs metric. The 5-percentile throughput is a measure of the cell edge throughput. Different points on the tradeoff curve between GAT and edge

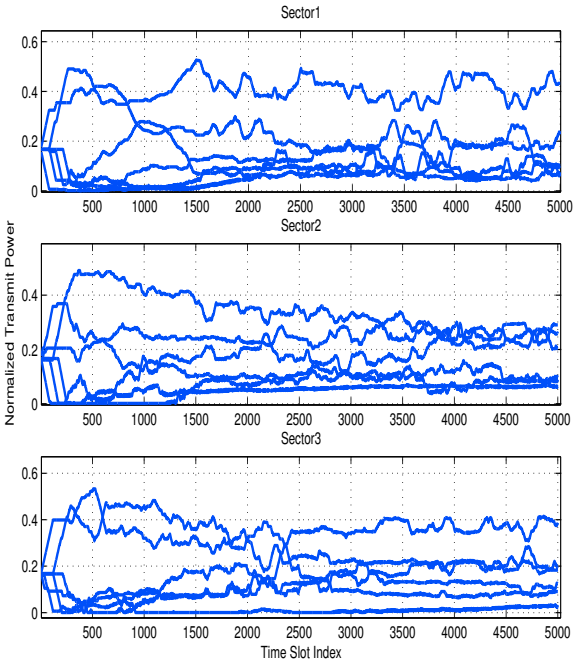

Fig. 1. Time series of normalized transmit powers on the different sub-bands for MGR

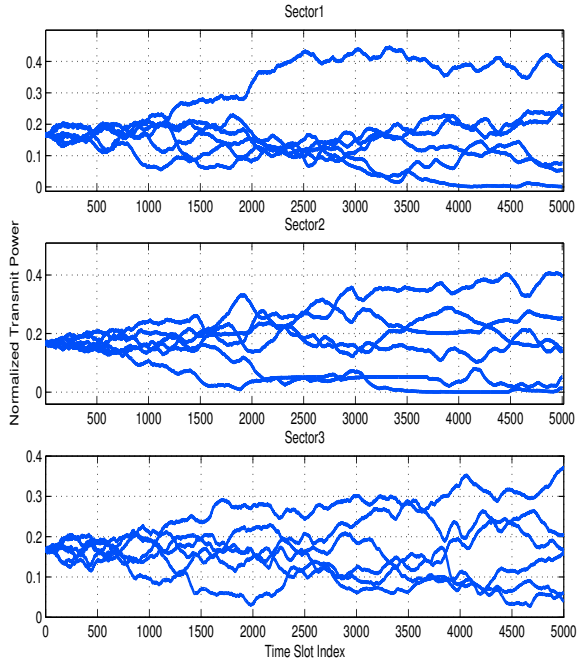

Fig. 2. Time series of normalized transmit powers on the different sub-bands for SA

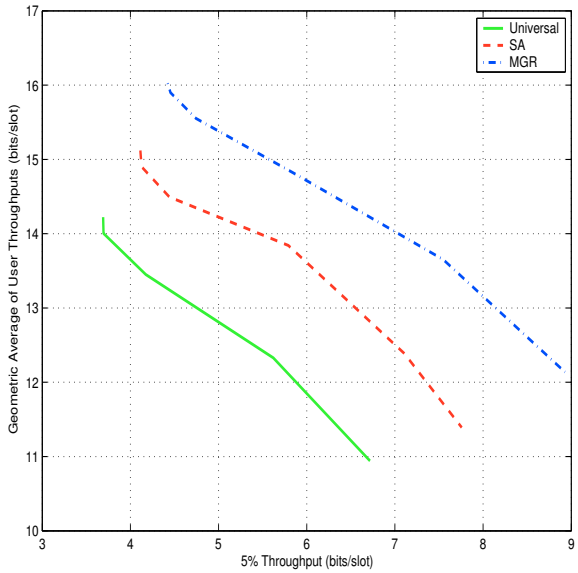

Fig. 3. Geometric average of user throughputs Vs. 5-% edge throughput. Uniform user distribution; fast fading.

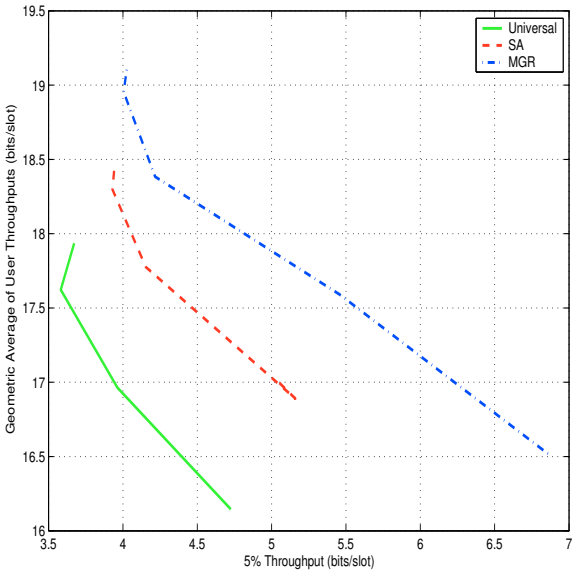

Fig. 4. Geometric average of user throughputs Vs. 5-% edge throughput. Non-uniform user distribution; fast fading.

throughput are obtained using the same scheduling algorithm but with different values of the minimum rate parameter. Two different scenarios, one where the users are distributed uniformly in each of the 57 sectors and another where the user distribution within each sector is non-uniform were simulated. Non-uniform user distribution is simulated as follows. User distribution for each sector is randomly chosen to be ”center” or ”edge” distribution. In the case of center distribution, the users are uniformly distributed in a region close to the base station and are guaranteed to have geometry (average SINR without fast fading) of greater than 6 dB. In the edge distribution, users are distributed uniformly in sector edges and have geometry below 0 dB.

Figure 3 shows the results for the case of uniform distribution of users. As can be seen from the figure, when the GAT is maintained at 12.8, the 5-percentile throughput can be increased by 34% using the SA algorithm and by 66% using the MGR algorithm relative to UNIVERSAL. Also observe that, as expected, the larger the 5-percentile throughput we want, the larger the gain in sector utility achieved by MGR and SA algorithms.

Figure 4 shows the results for the case of non-uniform distribution of users. The choice between the ”center” and ”edge” user distributions for each sector is kept the same across all algorithms and for all points along the curves. As can be seen from the figure, when GAT is maintained at 17.3, the 5-percentile throughput can be increased by 25% using the SA algorithm and by 55% using the MGR algorithm relative to the UNIVERSAL. It should be noted that using the proportional fair with minimum rate scheduling algorithm, increasing the minimum rate parameter further does not result in an increase in the edge throughput for the UNIVERSAL. Thus it is possible to achieve a much higher cell edge throughput using the MGR algorithm compared to UNIVERSAL.

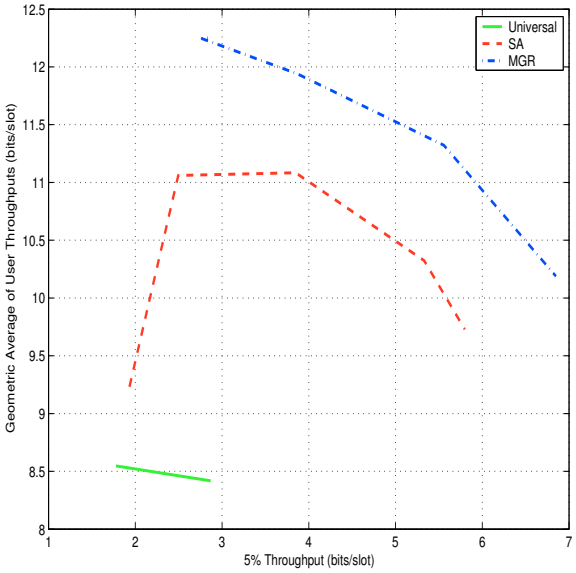

Fig. 5. Geometric average of user throughputs Vs. 5-% edge throughput. Uniform user distribution; NO fast fading.

Figure 5 shows the results for the case of uniform distribution of users, but *without* fast fading. Comparing these figures to those in Figure 3 shows that the gains of both algorithms are much larger in the case of no fast fading than with fading. (This is due to the fact that fast fading allows opportunistic schedulers, to a certain degree, avoid interference “automatically.” We discuss this phenomenon in some detail in [12].) We also see from the figure that the maximum cell edge throughput achievable by UNIVERSAL is substantially smaller compared to those of the SA and MGR algorithms. For the same 5-percentile throughput of about 3 bits/slot, the

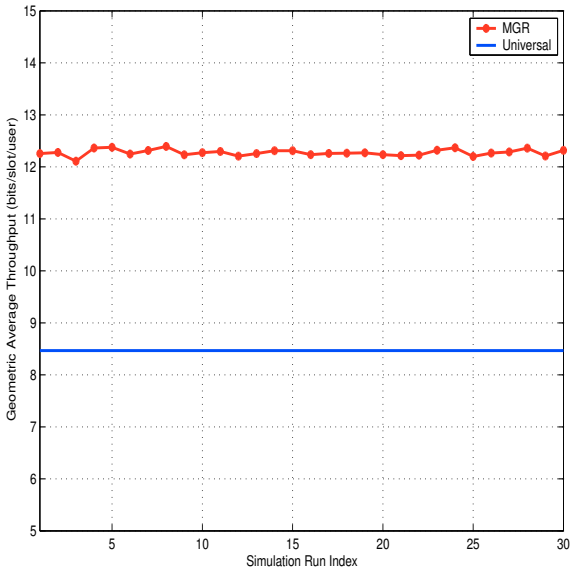

Fig. 6. Geometric average of user throughputs for different initial sub-band powers

GAT of MGR is 49% better than that of UNIVERSAL while that of SA is 35% better.

VIII. MGR Algorithm: Local Vs. Global Optimization

By its nature, MGR pursues a greedy *local* maximization of the system utility. Our simulations have shown that it gives significant performance improvement. However, a natural question is: For the optimization problem in this paper, how "bad" can a local maximum be compared to the global one? To shed some light into this, we study by simulation the performance of the MGR algorithm for different initial sub-band power settings. We generate thirty different initial power settings for the sub-band powers by randomly distributing the total sector power across the six sub-bands for each of the 57 sectors. We run the simulation for each of these initial power settings for the case of no fast fading and zero minimum-rate parameter. Figure 6 shows the GAT (our utility function) for the different runs together with the performance of UNIVERSAL (run with equal powers across the sub-bands). The deviation of the performance of the MGR algorithm for the different initial power settings is only about 4%, while the improvement over UNIVERSAL is about 44%. This suggests that, at least for the described setting, the local maxima of the problem have approximately equal values, and so local optimization leads to near optimality. This situation might be generic for FFR problems, although this certainly requires more research and evidence.

In addition, our experiment suggests that MGR is able to readjust sub-band power levels to near optimal values quickly from almost any initial power setting. This is important because it shows quick adaptivity to even dramatic changes in the system.

IX. Future work

Several avenues for future work are possible. We have focused on the forward link (base station to the users) in this paper. It is of interest to derive algorithms for the reverse link as well. Because of inherent asymmetries in the interference patterns forward link solutions may not carry over as is for the reverse link. We focused on best effort traffic in this paper while latency sensitive traffic was treated in our earlier work [11]. An overall scheme that combines these separate algorithms into a complete solution is another area for research. Finally, a study of fundamental limits on the performance gains from fractional frequency reuse, to benchmark the performance of the proposed algorithms, is of great interest.

References

[1] E. Altman, K. Avrachenkov, and A. Garnaev, "Closed form solutions for water-filling problem in optimization and game frameworks," in *Proceeding of INFOCOM'2008*, Phoenix, April 14-18, 2008.

[2] M. Andrews, L. Qian, A. L. Stolyar, "Optimal Utility Based Multi-User Throughput Allocation subject to Throughput Constraints," in *Proceeding of INFOCOM'2005*, Miami, March 13-17, 2005.

[3] T. Bonald, S. C. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *Proceedings of European Wireless Conference*, 2005.

[4] S. T. Chung, S. J. Kim, J. Lee, and J.M. Cioffi, "A game theoretic approach to power allocation in frequency-selective Gaussian interference channels," in *Proceedings of the IEEE International Symposium on Information Theory*, pp 316-316, July 2003.

[5] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proceedings of INFOCOM*, San Francisco, April 2003.

[6] S. Das and H. Viswanathan, "Interference mitigation through intelligent scheduling," in *Proceedings of the Asilomar Conference on Signals and Systems*, Asilomar, CA, November 2006.

[7] A. Gjendemsjo, D. Gesbert, G. E. Oien, and S. G. Kiani, "Optimal power allocation and scheduling for two-cell capacity maximization," in *Proceedings of the IEEE RAWNET (WiOpt)*, April 2006.

[8] B. Rengarajan, G. de Veciana, "Architecture and Abstractions for Environment and Traffic Aware System-Level Coordination of Wireless Networks: The Downlink Case," in *Proceedings of INFOCOM'2008*, Phoenix, April 14-18, 2008.

[9] A.L. Stolyar, "On the Asymptotic Optimality of the Gradient Scheduling Algorithm for Multi-User Throughput Allocation," *Operations Research*, 2005, Vol. 53, No.1, pp. 12-25.

[10] A. L. Stolyar, "Maximizing Queueing Network Utility subject to Stability: Greedy Primal-Dual Algorithm," *Queueing Systems*, 2005, Vol.50, No.4, pp.401-457.

[11] A. L. Stolyar, H. Viswanathan, "Self-organizing Dynamic Fractional Frequency Reuse in OFDMA Systems," in *Proceedings of INFOCOM'2008*, Phoenix, April 14-18, 2008.

[12] A. L. Stolyar, H. Viswanathan, "Self-organizing Dynamic Fractional Frequency Reuse Through Distributed Inter-cell Coordination: The Case of Best-Effort Traffic," *Bell Labs, Alcatel-Lucent, Technical Memo*, June 2008. http://cm.bell-labs.com/who/stolyar/be_dffr.pdf

[13] Third Generation Partnership Project 2, "Ultra Mobile Broadband Technical Specifications," http://www.3gpp2.org, March 2007.

[14] Third Generation Partnership Project, "Radio Access Network Work Group 1 Contributions," http://www.3gpp.org, September 2005.