A Semi-autonomous Algorithm for Self-organizing Dynamic Fractional Frequency Reuse on the Uplink of OFDMA Systems

Balaji Rengarajan
IMDEA Networks
Madrid, Spain
balaji.rengarajan@imdea.org

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

December 14, 2009

Abstract

Reverse link (or uplink) performance of cellular systems is becoming increasingly important with the emergence of new uplink-bandwidth intensive applications such as Video Share [13], where end users upload video clips captured through their mobile devices. In particular, it is important to design the system to provide good user throughput in most of coverage area, including at the cell edge. Soft fractional frequency reuse (FFR) is one of the techniques for mitigating inter-cell interference in cellular systems, leading to overall spectral efficiency enhancements and/or cell edge throughput improvements. We propose a novel algorithm that dynamically creates efficient soft FFR patterns on the uplink of orthogonal frequency division multiple access (OFDMA) based cellular systems; this allows the system to "automatically" adapt to user traffic distribution and system layout.

Our algorithm is based on systematically ascending towards a local maximum of the system-wide sum of user *utilities*, which depend on user throughputs. We show that this can be done in a semi-autonomous fashion: each sector does its resource allocation independently, with only an infrequent periodic exchange of *interference costs* between neighboring sectors. The proposed algorithm, called Multi-sector Gradient for Uplink (MGR-UL), allocates in-sector resources (power, frequency, time-slots to each user) in a way that simultaneously takes into account both the benefit to its "own" users' utility and the cost of creating interference to neighboring sectors; along with that each sector estimates the cost of interference to itself. Extensive simulation results show that significant performance benefits (up to 69% in total throughput in some typical scenarios) can be achieved with respect to a baseline approach. Simulations also show the automatic formation of soft FFR patterns.

1 Introduction

Fourth generation cellular systems, through their high data rates and system capacity, are expected to enable the true mobile broadband experience. Many new applications will emerge as a result of affordable broadband to mobile devices. In particular, we are likely to see the emergence of applications that are uplink-bandwidth intensive. An example of such an application is Video Share [13], where mobile users share video clips captured through their mobile devices while on the go. Thus it is imperative to improve uplink data throughput and spectral efficiency in addition to that of the downlink.

Cellular spectral efficiency is limited by, among other things, out-of-cell interference, with cell-edge user throughputs suffering the most. Mitigating out-of-cell interference is thus a promising approach to improving throughput performance, especially at cell edge. Traditionally, frequency reuse has been used to mitigate out-of-cell interference. However, traditional, "integer" frequency reuse (with each cell confined to a fixed part of available frequency spectrum) suffers from inefficient utilization of bandwidth. More recently, soft fraction frequency reuse (FFR) schemes have been proposed for orthogonal frequency division multiple access (OFDMA) systems, where each cell uses all available frequencies, but an intelligent choice of user transmit powers and frequency assignment is employed so as to minimize loss from interference across neighboring cells. In addition, it is desirable for an FFR scheme to be *dynamic* (or, *adaptive*), because an efficient FFR "pattern" is highly dependent on user spatial distribution and traffic demand (as well as on the system layout), which change with time. Finally, it is important for a dynamic FFR technique to be distributed (or, autonomous) "as much as possible", namely that different cells perform their resource allocation independently, with no or infrequent exchange of signaling information between neighbor cells. Our prior work towards the design of distributed dynamic FFR schemes [9, 10] focused on the downlink, and in this paper we propose an uplink algorithm.

Our proposed algorithm, called Multi-sector Gradient for Uplink (MGR-UL), systematically pursues local maximization of the system-wide sum of user *utilities*, which are functions of user throughputs. We show that this can be done in a *semi-autonomous* fashion: each sector does its resource allocation independently, with only an infrequent periodic exchange of *interference costs* between neighboring sectors. In-sector resource allocation procedure includes power, time-slot, and frequency assignment to the users; this procedure utilizes the interference costs (to neighbor sectors) as parameters; it also includes continuous estimation of the sector's "own" interference costs. Sectors send each other infrequent interference cost updates – this exchange is limited to neighbor sectors only. Thus, only a limited amount of information is exchanged between neighbor sectors, making practical implementation of the approach feasible.

Besides our earlier work on dynamic FFR for OFDMA downlink [9, 10], prior work on different aspects of resource allocation in OFDMA systems includes [7, 5, 3, 4, 1] – we refer reader to [9, 10] for a brief review. The concept of FFR for best effort traffic in the context of OFDMA systems has appeared in cellular network standardization technical contributions [11, 12] and in [6].

The paper is organized as follows. In Section 2 we describe the basic system model under consideration and the problem that is being addressed. In Section 3 we consider an idealized model of system behavior; for such model we formally derive the (asymptotically) optimal scheme, pursuing the fastest system utility ascend. Section 4, first, outlines issues that need to be addressed by a resource allocation scheme in a real system, and then describes the MGR-UL algorithm (motivated by the

idealized algorithm of Section 3) designed for practical implementation. In Section 5 we compare the performance of the proposed algorithm with that of the optimal algorithm (for idealized model) for a simple two sector system. The simulation studies, comparing the performances of MGR-UL and a baseline algorithm in a realistic setting are given in Section 6. We conclude with a summary and discussion of future work in Section 7.

2 Basic model and problem statement

We consider the uplink (from mobile user to base station) of a multi-sector OFDMA system, where each user i is assigned to one of the sectors m; $\mathcal{I}$ and $\mathcal{M}$ are the finite sets of users and sectors, respectively. (Notation $i \in m$ will mean "user i is within sector m".) Frequency band is divided into J equal size subbands, indexed by j; $\mathcal{J} = \{1, \ldots, J\}$ is the set of subbands. Each subband consists of C resource blocks, each of bandwidth W. N_0 is spectral noise density. Our assumptions about propagation gains will be specified later.

The system operates in discrete time, over time-slots $t = 0, 1, \ldots$. In each time-slot each sector schedules a subset of "its" users to transmit, and for each scheduled user it assigns: the subbands, the number of resource blocks in each subband, and the power-per-resource-block (and per-subband) to use. The total power assigned to any user in any time-slot cannot exceed P^*.

A key feature of an OFDMA system is that transmissions of different users *within a sector* do not interfere with each other (because they use different frequencies), while transmissions of users within different sectors do interfere with each other *if they happen to use same frequencies*. This in particular means that two transmissions within same sector or using different subbands (within same sector or not) never interfere with each other. Two transmissions using the same subband in different sectors interfere with each other whenever the resource blocks used for the transmissions are overlapping in frequency. If different resource blocks within the same subband are used, then the transmissions in the different sectors do not interfere with each other.

In this paper, we consider best-effort user traffic. Each user i has an associated (concave, smooth, strictly increasing) utility function $U_i(X_i)$ of its average (over time) achieved rate X_i. ($U_i(X_i) = \log(X_i)$ for a proportional fair objective.) The goal is find a scheduling strategy so as to maximize the total utility of the system, $\sum_i U_i(X_i)$. Note that here "scheduling" strategy is understood broadly, because it includes not only the choice of users to transmit, but also their subband, resource block and power assignments.

3 An idealized model of system-wide utility maximization

To motivate our proposed scheme (given in Section 4), we first consider a further idealized model, for which we can formally derive a scheme that performs, essentially, a gradient ascend towards a local maximum of the system-wide utility. The key additional idealization is, roughly speaking, to assume that a transmission in a time slot experiences the interference that is a time-averaged interference (over a certain window), as opposed to actual instantaneous one. This assumption is well suited for modeling a system where sectors do not (or cannot) have good estimates of actual *instantaneous* interference levels, and thus have to employ an autonomous or semi-autonomous control; this is the

situation we are interested in the present paper. (If we were interested in a multi-sector system that somehow is controlled by a central entity that knows all instantaneous propagation gains, then, at least in principle, the inter-sector coordination can be done on a slot-by-slot basis, and from a purely scheduling point of view such system is equivalent to a single-sector one.)

We will also assume that interference in a subband is spread uniformly over all resource blocks within the subband. (This assumption is actually quite realistic, if logical resource blocks use frequency hopping over physical frequency tones (subcarriers) within a subband.) For each sector, we assume that the set of its scheduling choices is finite (although it can be very large). Finally, we assume that the propagation gain G_i^m (path loss and shadowing) from user i to (base station of) sector m is constant.

3.1 Definition of system utility

Suppose that in each time slot, each sector m has finite set S_m of possible choices (decisions) of how to allocate resource blocks and transmit powers to its users $i \in m$. Associated with each choice $s \in S_m$, are the number of resource blocks $h_{ij}(s)$ and power-per-resource-block $p_{ij}(s)$, allocated to each user $i \in m$ in each subband j. Naturally, for each s, the total power allocated to each user i cannot exceed maximum available mobile power, $\sum_j h_{ij}(s)p_{ij}(s) \le P^*$, and the number of allocated resource blocks within each subband j cannot exceed C: $\sum_i h_{ij}(s) \le C$. (Each discrete finite set S_m can be very large, for example obtained by quantizing the "true", continuous set of all possible allocation choices; the cardinality of set S_m is irrelevant for the purposes of this section.)

Suppose each sector m does scheduling in a way such that decision $s \in S_m$ is picked in the fraction ϕ_s^m of all time-slots, $\sum_s \phi_s^m = 1$. Denote $\phi = \{\phi_s^m,\ s \in S_m, m \in \mathcal{M}\}$.

Given ϕ, the system utility U is defined as follows:

$$U = \sum_m \sum_{i \in m} U_i(x_i),$$

where each U_i is strictly increasing continuously differentiable function (e.g. log) of user i average achieved rate

$$x_i = \sum_{s \in S_m, m \ni i} \phi_s^m \sum_j h_{ij}(s) r_{ij}(s), \tag{1}$$

where user i rate per-resource-block in subband j is

$$r_{ij}(s) = W \log_2 \left(1 + \frac{G_i^m p_{ij}(s)}{N_0 W + I_j^m / C} \right), \tag{2}$$

and, finally, the average interference in subband j of sector m is

$$I_j^m = \sum_{k \neq m} \sum_{s \in S_k} \phi_s^k \sum_{i \in k} G_i^m h_{ij}(s) p_{ij}(s). \tag{3}$$

Remark. The above definition of utility specifies our main idealized assumption: the interference levels I_j^m are constant, equal to their average values (if scheduling choices are made according to given ϕ).

Clearly, the utility U is a continuously differentiable function of ϕ (in the domain of ϕ where all $x_i > 0$). It will be convenient, however, to view it as $U(\phi) = V(\phi, I(\phi))$, where the "intermediate" function $V(\phi, I)$ is a continuously differentiable scalar function of ϕ and $I = \{I_j^m, m \in \mathcal{M}, j \in \mathcal{J}\}$, defined by $V(\phi, I) = \sum_m \sum_{i \in m} U_i(x_i)$, (1) and (2), and $I(\phi)$ is continuously differentiable vector-function of ϕ, defined by (3).

Also, it will be convenient to rewrite (2) as

$$r_{ij}(s) = \frac{W}{\ln 2} \ln\left(1 + f_{ij}(s)\right), \tag{4}$$

where

$$f_{ij}(s) = \frac{G_i^m p_{ij}(s)}{N_0 W + I_j^m / C}.$$

We get the following expressions for partial derivatives:

$$\frac{\partial V}{\partial \phi_s^m} = \sum_{i \in m} U_i'(x_i) \sum_j h_{ij}(s) r_{ij}(s), \tag{5}$$

$$a_j^m \doteq -\frac{\partial V}{\partial I_j^m} = \sum_{i \in m} U_i'(x_i) \sum_{s \in S_m} \phi_s^m h_{ij}(s) \frac{W [f_{ij}(s)]^2}{C(\ln 2)(1 + f_{ij}(s)) G_i^m p_{ij}(s)}, \tag{6}$$

$$\frac{\partial I_j^m}{\partial \phi_s^k} = 0, \quad k = m,$$

$$\frac{\partial I_j^m}{\partial \phi_s^k} = \sum_{i \in k} G_i^m h_{ij}(s) p_{ij}(s), \quad k \neq m. \tag{7}$$

3.2 Dynamics of system utility

Consider now the following dynamic system, which evolves in discrete time $t = 0, 1, 2, \ldots$; t is a time-slot index. Let a fixed set (of fractions) $\phi(t)$ be the system state at time t, and $U(\phi(t))$ is utility at t. At time $t + 1$ each sector m can choose one allocation $s_m(t + 1) \in S_m$; when such choice is made, the state changes to $\phi(t + 1)$, as follows:

$$\phi_s^m(t + 1) = \eta \cdot 1 + (1 - \eta)\phi_s^m(t), \quad s = s_m(t + 1),$$

$$\phi_s^m(t + 1) = (1 - \eta)\phi_s^m(t), \quad s \neq s_m(t + 1),$$

where $\eta > 0$ is a small parameter. (The interpretation of such state evolution is as follows: each frequency $\phi_s^m(t)$ is the (exponentially) averaged frequency of choosing decision s in the past.) Correspondingly, the system utility becomes $U(\phi(t+1))$. (Here again we used the idealization that only the current average values of the interferences affect the transmission rates, not the instantaneous interference levels within each slot – these depend on the actual scheduling choices $s_m(t + 1)$ made by the sectors.)

Now, since function V is continuously differentiable (and η is small), we can write:

$$U(\phi(t + 1)) - U(\phi(t)) = \eta \sum_m \sum_{i \in m} U_i'(x_i)(\sum_j h_{ij}(s_m(t + 1)) r_{ij}(s_m(t + 1)) - x_i) - \tag{8}$$

$$-\eta \sum_m \sum_{k\neq m} \sum_{i\in m} \sum_j a_j^m \left(G_i^k h_{ij}(s_m(t+1)) p_{ij}(s_m(t+1)) - \sum_{s\in S_m} \phi_s^m G_i^k h_{ij}(s) p_{ij}(s) \right) + o(\eta). \qquad (9)$$

Therefore, to maximize the increment of utility from slot t to slot $t+1$ (up to $o(\eta)$ error, i.e. to maximize the scalar product of $\phi(t+1)) - \phi(t)$ and the gradient of U at $\phi(t)$), each sector m can *independently* choose allocation

$$s_m(t+1) \in \arg\max_{s\in S_m} \left(\sum_{i\in m} U_i'(x_i) (\sum_j h_{ij}(s) r_{ij}(s)) \right) - \left(\sum_{k\neq m} \sum_j a_j^k \sum_{i\in m} G_i^k h_{ij}(s) p_{ij}(s) \right). \qquad (10)$$

The first term is interpreted as the utility gain in sector m due to allocation s, and the second term is the utility loss in other sectors due to interference caused by allocation s. We emphasize that, as long as the current interference costs a_j^k and gains G_i^k are "known" to sector m, the optimization is "local" to the sector.

Thus, if the system dynamics is governed by scheduling rule (10), the state $\phi(t)$ is driven towards a local maximum of function $V(\phi)$.

Remark. If we drop the second, cost term in rule (10), we obtain the well known Gradient scheduling rule (see [8] and references therein), which is known to be optimal for a single-sector system (or a centrally controlled multi-sector system). This name is appropriate for the more general rule (10) as well, because, as we saw, it also uses the gradient of utility function, except now it is a system-wide utility of a multi-cell network.

3.3 Approach to utility maximization in a real (not idealized) system

The idealized model of this section and its dynamics motivate and substantiate the following approach to system-wide utility maximization in a real multi-cell system. It consists of two parts:

1. Each sector m continuously estimates and updates its current interference costs a_j^m, i.e. the sensitivity of its utility to the interference level in each subband j. This is done by time-averaging (performed along with in-sector scheduling), which maintains an approximation of expression (6). These interference costs a_j^m are periodically communicated to other sectors (those neighbor sectors k that create non-negligible interference to m), which use them for making scheduling decisions (see item 2).

2. Each sector m makes a scheduling decisions, which approximate the rule (10). (Implementing rule (10) exactly is not feasible in most cases, due to computational complexity.) Note that rule (10) uses the interference costs a_j^k in neighboring sectors, periodically received from those sectors.

Thus, the approach attempts is to constantly "drive" (using, essentially, a gradient estimate) the system towards an operating point where a local maximum of utility is attained. A specific scheme following this approch is given in Section 4. As we will see in the simulation results of Section 6, the scheme indeed produces a substantial improvement in system utility.

4 Resource allocation scheme for a real system

4.1 Definitions, assumptions and constraints pertaining to a real system

Any scheme for a real system has to take into account the following features/constraints.

1. Interference experienced by a transmission in a given time slot and given resource block is not the time-average interference, but really is instantaneous, depending (among other things) on instantaneous user/resource-block/power scheduling assignment in neighboring sectors.

2. Transmission of a data packet may not be completed within one time slot, due to insufficient instantaneous signal to interference ratio. Several HARQ retransmissions may be then required.

3. To allow for HARQ feedback, time slots are divided into several "interlaces", say 5. This means that if a packet is transmitted in slot t and requires retransmission(s), they are scheduled in slots $t+5, t+10$, and so on. This in particular means that not all resource blocks may available for scheduling new transmissions in a given time slot – some may be already taken by retransmissions.

4. Retransmissions of a packet may not occur in the same physical resource block (same frequency) as original transmission, because a frequency hopping may be employed.

For a real system, notation G_i^ℓ will be used for the *average* propagation gain (path loss and shadowing) from user i to (base station of) sector ℓ. The average propagation gains can be obtained from downlink pilot strength measurements made at the mobile and fed back to the base station. (Such measurements are available, because they are required to enable handoffs.) The average propagation gain on the downlink will be the same as that on the uplink because of channel reciprocity for path loss and shadow fading. Fast fading, which is the only component of the gain that is different on the two links, is averaged out.

When we talk about resource allocation in a given sector, it is usually denoted by m.

4.2 Interference costs

Non-negative cost a_j^m is the cost to the utility of sector m of unit interference increase in subband j. It is a dynamic quantity computed and maintained by each sector m as described below. It is changing slowly with time, and each sector sends periodic (infrequent) updates of its costs a_j^m to all other sectors ℓ. (In reality it sends it only to the neighbor sectors, those that cause sufficiently high interference to m.)

4.3 Scheduling objective

In each time slot, for each of its users $i \in m$ in each subband j, sector m needs to choose the number of resource blocks H_{ij} to be assigned and the transmit power P_{ij} per resource block, such that

$$\sum_j H_{ij}P_{ij} \le P^* \quad \text{and} \quad \sum_j H_{ij} \le C, \quad \forall i.$$

The objective is to maximize the quantity

$$\left(\sum_{i\in m} U_i'(X_i)(\sum_j H_{ij}R_{ij})\right) - \left(\sum_{k\neq m}\sum_j a_j^k \sum_{i\in m} G_i^k H_{ij}P_{ij}\right), \tag{11}$$

where X_i is the current average rate of user i (updated as described below), R_{ij} is the rate per-resource-block

$$\begin{aligned} R_{ij} &= W\log_2\left(1+\frac{G_i^m P_{ij}}{N_0 W + I_j^m/C}\right) \\ &= \frac{W}{\ln 2}\ln\left(1+F_{ij}\right), \end{aligned} \tag{12}$$

I_j^m is the average interference power to sector m in the entire subband j, and we denoted the user i SINR by

$$F_{ij} \doteq \frac{G_i^m P_{ij}}{N_0 W + I_j^m/C}.$$

4.4 Scheduling algorithm: Multi-cell Gradient for Uplink (MGR-UL)

The computational complexity of maximizing (11) is typically prohibitively high. Therefore, we propose a heuristic algorithm, consisting of the following steps, performed in each time-slot.

Step 1. We compute the transmission power P_{ij} per resource block, which is to be used *if user i is actually scheduled in subband j*. This power P_{ij} is determined so that it maximizes the utility value minus cost,

$$U_i'(X_i)R_{ij}(P_{ij}) - \sum_{\ell\neq m} a_j^\ell G_i^\ell P_{ij},$$

subject to being within $[0, P^*]$ and not causing more than the interference $\bar{P}$ per resource block to any other sector $\ell \neq m$. This gives

$$P_{ij} = \max\left(\min\left(Y, P^*, \min_{\ell\neq m}\frac{\bar{P}}{G_i^\ell}\right), 0\right)$$

where

$$Y = \frac{U_i'(X_i)W}{(\ln 2)\sum_{\ell\neq m} b_j^\ell G_i^\ell} - \frac{1}{\hat{F}_{ij}}$$

and

$$\hat{F}_{ij} = \frac{G_i^m}{N_0 W + I_j^m/C}.$$

Estimates of the average gains G_i^ℓ to neighbors are obtained from pilot signal measurements as explained in Section 4.1. The estimate of average "gain-to-interference-plus-noise" ratio $\hat{F}_{ij}$ is obtained by averaging (see Section 4.5) its instantaneous values $\hat{\Gamma}_{ij}$. (In turn, $\hat{\Gamma}_{ij}$ is obtained by measuring received pilot SINR for user i in subband j. The specific way we use in simulations is described in Section 6.1.3.)

Step 2. For the purposes of scheduling, the transmission rates are assumed to be

$$\hat{R}_{ij}(P_{ij}) = W \log_2 \left(1 + \hat{\Gamma}_{ij} P_{ij}\right).$$

Step 3. The packet size per-resource-block is chosen, based on $\hat{R}_{ij}(P_{ij})$; we chose it to be just equal to $\hat{R}_{ij}(P_{ij})$.

Step 4 (ACTUAL SCHEDULING). We pick a resource block at random among all still available resource blocks across all subbands, suppose it happens to be in subband j. In this resource block we schedule user i for which the value of

$$Z_{ij} = U_i'(X_i)\hat{R}_{ij}(P_{ij}) - \sum_{\ell \neq m} a_j^\ell G_i^\ell P_{ij},$$

is maximal among those users i which still have at least κP_{ij} "leftover" power to transmit, where $\kappa \in [0,1]$ is a parameter; we chose $\kappa = 0.9$. (The leftover power for the user is updated after it is scheduled in a resource block.) Then we pick another (unused yet) resource block at random, and so on, until we make allocation (perhaps, null) in each resource block.

Step 5 (OPTIONAL). For each allocated resource block and the corresponding user, we check if the scheduling objective can be improved by reallocating the user to a still available resource block in a different subband; if so, we make the "best" such reallocation. This step is performed for each of the resource blocks, allocated in Step 1. (In our simulations this step is moot, because the setting is such that all resource blocks are used in Step 4.)

4.5 $\hat{F}_{ij}$ updates

$\hat{F}_{ij}$ is updated once per scheduling interval for all (ij):

$$\hat{F}_{ij} := (1 - \beta_0)\hat{F}_{ij} + \beta_0 \hat{\Gamma}_{ij},$$

where $\beta_0 > 0$ is a (small) averaging parameter.

4.6 Average rate X_i updates

X_i is updated once per scheduling interval for all users i:

$$X_i := \beta \overline{R}_i + (1 - \beta)X_i, \tag{13}$$

where $\beta > 0$ is a (small) averaging parameter and $\overline{R}_i$ is the total size of all packets of user i whose transmission was *successfully completed* in this slot. If user i was not transmitting or did not complete any transmission, then $\overline{R}_i = 0$.

4.7 Calculation of costs a_j^m

As we allocate resource blocks during scheduling, and along with that, we update costs a_j^m as follows. If a resource block is assigned to user i (whether it is newly assigned, or user i HARQ transmission

continues from the previous slot), we do

$$a_j^m := \beta_1 \frac{U_i'(X_i) W [F_{ij}]^2}{(\ln 2)(1 + F_{ij}) G_i^m P_{ij}} + (1 - \beta_1) a_j^m, \tag{14}$$

where $F_{ij} = \hat{F}_{ij} P_{ij}$;
for an "unused" resource block (in subband j), we do

$$a_j^m := (1 - \beta_1) a_j^m. \tag{15}$$

4.8 Average rate X_i initialization

In reality, the set of users in a sector changes with time. In addition, users that are active may not have data to send all the time. This makes the issue of "good" initial value of X_i important. (Although, it does not arise in the simulations presented in this paper, where we assume all users always have data to send.) One option can be as follows.

When a new user or data session "arrives," its X_i is initially set to some, relatively low value X_{init}, a parameter. (It may depend on the user priority class, etc.) In the time slots when this user has no data to send, its X_i is updated so that it "drifts" towards X_{init}:

$$X_i := \beta X_{init} + (1 - \beta) X_i.$$

Thus, if user i traffic flow consists of small chunks of data, then marginal utility $U_i'(X_i)$ (of one bit of user i data) is roughly kept constant at (a relatively high level) $U_i'(X_{init})$, regardless of the user's channel quality, which is a reasonable thing to do in this case.

On the other hand, if user i, for example, uploads a large file for a significant period of time, its marginal utility will be $U_i'(X_i)$, that is it will depend on the actual average rate X_i observed by the user. (Again, in our simulations, all users are like that.)

5 Simple two-sector system

Before we show simulation results for a 19-cell, 57-sector system, we present some numerical and simulation results for a simple one-dimensional, symmetric two-sector system. The reason is that for such a system, the corresponding idealized model can be optimized numerically, and so we get a sense of how well our algorithm, MGR-UL, performs compared to an *optimal* FFR scheme. (We note that, formally speaking, there is no guarantee that the performance of the optimal scheme for an idealized system will be better than that of any realistic scheme in a real system. It is natural to expect, however, that realistic schemes not involving coordination of transmissions between sectors on a slot-by-slot basis, and thus making scheduling decisions based on observed average interference levels, will typically perform worse than an optimal idealized system.)

Consider the one dimensional system shown in Fig. 1. There are two cells, 1 and 2, with corresponding base stations BS 1 and BS 2. The frequency spectrum is divided into two subbands, 1 and 2. Each cell has one sector, and so in this section we will use terms 'cell' and 'sector' interchangeably. Each cell serves two users. The users in cell 1 are located at points u_1 (close to

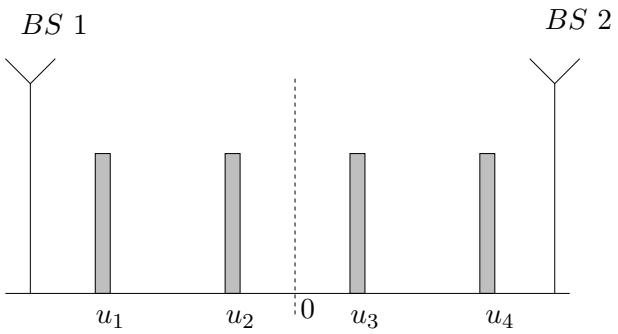

Figure 1: Symmetric one-dimensional system.

BS 1) and u_2 (closer to point 0, the boundary between the cells). The location of BS 2 and the distribution of users in cell 2 are symmetric to those in cell 1 with respect to the boundary between the cells. We call the users located at u_1 and u_4 'near' users, and those at u_2 and u_3 the 'far' users. The base stations are located 2 Km apart, and the propagation gain is modeled using a log distance path loss model with path loss exponent 3. The maximum transmit power of a user, denoted by P^*, is set to be such that the received power from the cell edge (point 0) is 10 times the total noise power, which is the noise spectral density N_0 times the total system bandwidth. The rate at which a user is served is determined using Shannon's formula (as specified in previous sections), but with SINR capped at 25 dB. The utility function of user i is $\log X_i$.

We obtain the performance of the proposed MGR-UL scheme, assuming one resource block in each subband, and running the algorithm as specified in Section 4.

Then we consider the idealized model, as described in Section 3. For this model, we further assume that each subband is divided into *infinite* number of resource blocks. Given this, a power/resource block allocation decision s is determined by:
$\psi_{ij} \in [0, 1]$, the fraction of all resource blocks in subband j allocated to user i (this is a "normalized version" of h_{ij});
$\pi_{ij} \in [0, P^*]$, the power density, allocated to user i in subband j (this is a "normalized version" of p_{ij}).
It is easy to observe that, since fractions ψ_{ij} and densities π_{ij} can be real numbers, the utility of the idealized system is maximized if each sector chooses one *fixed* optimal allocation s_0. (In other words, speaking informally, for each sector m, the set of fractions ϕ^m is such that $\phi^m_{s_0} = 1$ and $\phi^m_s = 0$ for all other possible allocations s.) We numerically find an optimal *symmetric* allocation, namely such that:
$\pi_{1,1} = \pi_{4,2}$, $\pi_{1,2} = \pi_{4,1}$, $\pi_{2,1} = \pi_{3,2}$, $\pi_{2,2} = \pi_{3,1}$,
$\psi_{1,1} = \psi_{4,2} = 1 - \psi_{2,1} = 1 - \psi_{3,2}$,
$\psi_{1,2} = \psi_{4,1} = 1 - \psi_{2,2} = 1 - \psi_{3,1}$.
The optimization is over 6 variables, and its done by exhaustive search over a discretized sets of possible values of each variable.

In addition, to quantify the advantage of FFR over standard reuse schemes, we consider two variants of the idealized model that correspond to full (universal) frequency reuse and 1/2 reuse.

Namely, for the full reuse, where all subbands are used uniformly, without any preference, we impose the following additional optimization constraints: $\pi_{i1} = \pi_{i2}$ and $\psi_{i1} = \psi_{i2}$ for all users i.

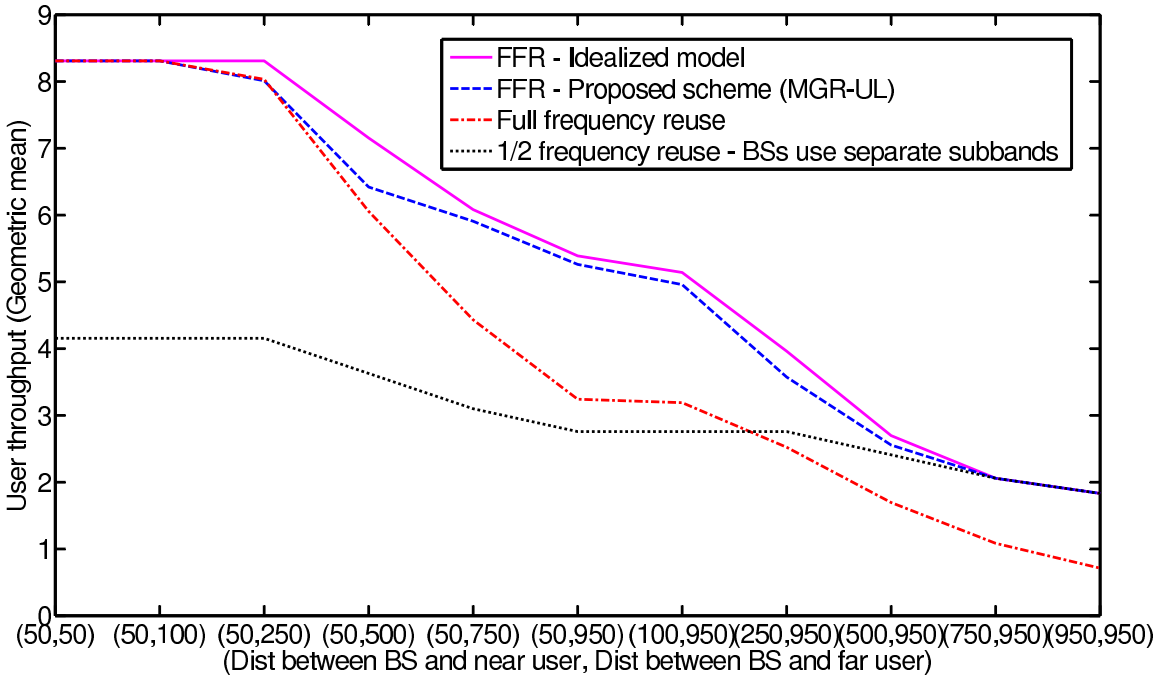

Figure 2: Performance of FFR schemes as a function of user location.

The second variant corresponds to 1/2 reuse. Here cell 1 can only use subband 1, and cell 2 can only use subband 2. Therefore, the additional constraints in this case are:
$\pi_{i,2} = 0$ and $\psi_{i,2} = 0$ for users $i = 1, 2$ in cell 1; $\pi_{i,1} = 0$ and $\psi_{i,1} = 0$ for users $i = 3, 4$ in cell 2.

Fig. 2 plots the geometric mean of the achieved user throughputs, normalized by the number of users (four) as the locations of the users are varied. In the leftmost scenario plotted along the x-axis, both users in the cells are close to their respective base stations, and the impact of interference is very small. In this case, reusing the entire frequency spectrum in both cells result in the best performance and the 1/2 reuse scheme (that tries to sacrifice bandwidth to mitigate interference) performs very poorly. Note that both the (optimal) idealized FFR scheme and the proposed MGR-UL scheme also result in performance virtually identical to that of full frequency reuse. In the rightmost scenarios, all the users are close to the cell boundary, and the rate of transmissions in sub-bands that are used in both cells are affected significantly by interference. In this case, the 1/2 frequency reuse is the best policy, while full reuse results in significant degradation in performance. Both the idealized FFR scheme and the MGR-UL scheme also "reduce" to the 1/2 reuse policy in such a situation. In the intermediate scenarios plotted on the x-axis, one of the users in each cell is close to the base station, and the other is close to the cell boundary. Clearly, neither full frequency reuse nor 1/2 reuse are close to the optimal idealized FFR policy in such a scenario, which is the type of scenario where the benefit of fractional frequency reuse is largest. Also, the proposed MGR-UL scheme achieves throughputs that are very close to the optimal idealized FFR policy. The important observation here is that the dynamic MGR-UL scheme achieves throughputs very close to the optimal idealized FFR policy under *all* the tested scenarios, demonstrating the efficacy and adaptability of the proposed method.

Parameter	Assumption
Cell Layout	Hexagonal 57 sector
Inter-site distance	856 m
Path Loss Model	$L = 133.6 + 35 \log_{10}(d)$
Shadowing	Log Normal with 4.2 dB Std. Dev.
Penetration Loss	10 dB
Noise Bandwidth	0.47 Mhz
Mobile Tx Power	23 dBm
BS Antenna Gain	17 dB
Mobile Antenna Gain	1 dB
BS Noise Figure	3 dB
Channel Model	No fading, Frequency-selective fading equivalent

Table 1: Propagation parameter values used in the simulation results

6 Simulations

6.1 System model for simulations and MGR-UL algorithm implementation aspects.

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.

Standard propagation parameters as listed in Table 1 are used to determine the received signal power level for a given transmit power level. For these parameters, with the site-to-site distance set at 856m, the cell edge SNR (signal to thermal noise ratio, when there is no interference from surrounding cells, assuming total available power is distributed uniformly over the entire bandwidth) turns out to be 22.4 dB. A small cell size, typical of urban morphology, has been chosen since gains are generally larger with smaller cells because of higher interference levels.

To demonstrate the effect of fast fading we run the simulations with and without it. When fast fading is used in the simulations, the model is representative of frequency-selective Rayleigh fading with temporal characteristics captured through Jakes fading model with vehicle speed of 20 Km/hr and carrier frequency of 2 Ghz. The frequency-selectivity is modeled by simulating independent fading across sets of coherence bands. In our simulations we consider 6 sub-bands that are divided into three sets of coherence bands each with two sub-bands.

6.1.1 Traffic model

The full buffer traffic model is used for all the simulation experiments. In this case, the assumption is that all users have an infinite amount of back-logged traffic. The simulations are run for 5000 slots, each of 1 ms duration.

6.1.2 Frequency hopping

In our simulation, we consider an OFDMA system with 18 frequency resource blocks divided into 6 sub-bands, 3 resource blocks in each. Random frequency hopping is implemented from slot to slot by permuting the resource block indices (within a sub-band) independently across the different sub-bands and sectors.

6.1.3 SINR estimation at the base station

Instantaneous pilot SINR $\hat{\Gamma}_{ij}$ of the mobiles needs to estimated at the base station for the purposes of scheduling and transmit power allocation, described in Section 4.4. The "true instantaneous" value of pilot SINR calculated at the base station is

$$\tilde{\Gamma}_{ij} = \frac{\tilde{G}_i^m}{N_0 W + \tilde{I}_j^m / C}, \tag{16}$$

where $\tilde{G}_i^m$ is the instantaneous gain to the serving sector m, and $\tilde{I}_j^m$ is instantaneous interference in subband j (normalized so that it is interference per-subband). Clearly, $\tilde{\Gamma}_{ij}$ is highly variable, in particular because it depends on the instantaneous gains of neighbor-sector users (to sector m) and their instantaneous power and resource block allocation.

Since there is a delay in utilizing the calculated SINR for packet scheduling purposes, it is more appropriate to use $\hat{\Gamma}_{ij}$ which is a short-term average of the values of $\tilde{\Gamma}_{ij}$. Moreover, reasonable way to do such averaging is to actually average the rates corresponding (by Shannon formula) to the instantaneous SINR, and then convert the average rate back to the SINR form. Namely, in the simulations we do the following (averaging) update within each time slot:

$$\tilde{R}_{ij} := \beta_1 W \log_2 \left(1 + \tilde{\Gamma}_{ij}\right) + (1 - \beta_1)\tilde{R}_{ij}, \tag{17}$$

(we chose $\beta_1 = 1/5$), and then determine $\hat{\Gamma}_{ij}$ from the equation

$$\tilde{R}_{ij} = W \log_2 \left(1 + \hat{\Gamma}_{ij}\right). \tag{18}$$

6.1.4 Mobile transmit power and actual scheduling of transmissions

The transmit power for each mobile in each sub-band is determined by the algorithm in Section 4.4 for the MGR-UL algorithm. For the baseline scheme the transmit powers are set according to the description in Section 6.2

In our simulations we, in fact, use a generalization of the Gradient scheduling algorithm (see [2]) which allows us to introduce minimum rate requirements of the form $X_i \geq R_{min}$ for some constant $R_{min} \geq 0$. The generalized algorithm maintains a *token counter* variable T_i (updates as described in [2]) for each user i, and uses $\exp(aT_i)U_i'(X_i)$ in place of $U_i'(X_i)$ wherever the latter appears in the MGR-UL description given in Section 4.4. (When all minimum rate requirements are zero, the algorithm reverts to the exact form of Section 4.4.)

Step 5 of MGR-UL is not used in our simulations, since we have more users than resource blocks, and therefore all resource blocks are "taken" in Step 4.

6.1.5 Rate computation for actually scheduled transmissions

For all scheduled users in a given slot, the number of successfully received bits during the slot (which may be less than the packet size) needs to be computed. If, for example, a user i transmission uses multiple resource blocks of subband j, then the actual rate "in each resource block" d is computed as

$$W \log_2 \left(1 + \tilde{\Gamma}_{ij,d} P_{ij}\right),$$

where $\tilde{\Gamma}_{ij,d} P_{ij}$ is the actual SINR in d. From this rate and the number of OFDM symbols per slot, 8 in the simulations, the actual number of bits received is computed.

6.2 Baseline algorithm

As a baseline against which we evaluate the performance of MGR-UL, we consider the following algorithm. (It will be referred to as 'Baseline'.) All other aspects of the system model are the same as in the previous sub-section.

Transmission power setting. Each sector m maintains a variable SINR target $\hat{F}_m$. The (potential, if scheduled) power P_{ij} per resource block for user i within *any* subband j (all subbands are "treated" the same way) is calculated to achieve SINR $\hat{F}_m$, or is set to the maximum power P^* if $\hat{F}_m$ is not achievable.

Scheduling. After the powers per resource block P_{ij} are calculated, scheduling within a time slot is done in exactly same way as for MGR-UL, except without taking into account the interference costs; namely, with $Z_{ij} = U_i'(X_i)\hat{R}_{ij}(P_{ij})$. (The average rate and pilot SINR updates are done the same way as well.)

SINR target $\hat{F}_m$ updates. Each sector k measures the average (over time) Interference-over-Thermal-noise level I_{oT}^k. If I_{oT}^k exceeds some maximum level (parameter) I_{oT}^* in a time slot, sector k sends an "overload message" (one bit) to is neighboring sectors m indicating such event. In every time slot, sector m increases its $\hat{F}_m$ by $\delta_1 > 0$, unless it receives an overload message from at least one of its neighbors – then it decreases $\hat{F}_m$ by $\delta_2 > 0$.

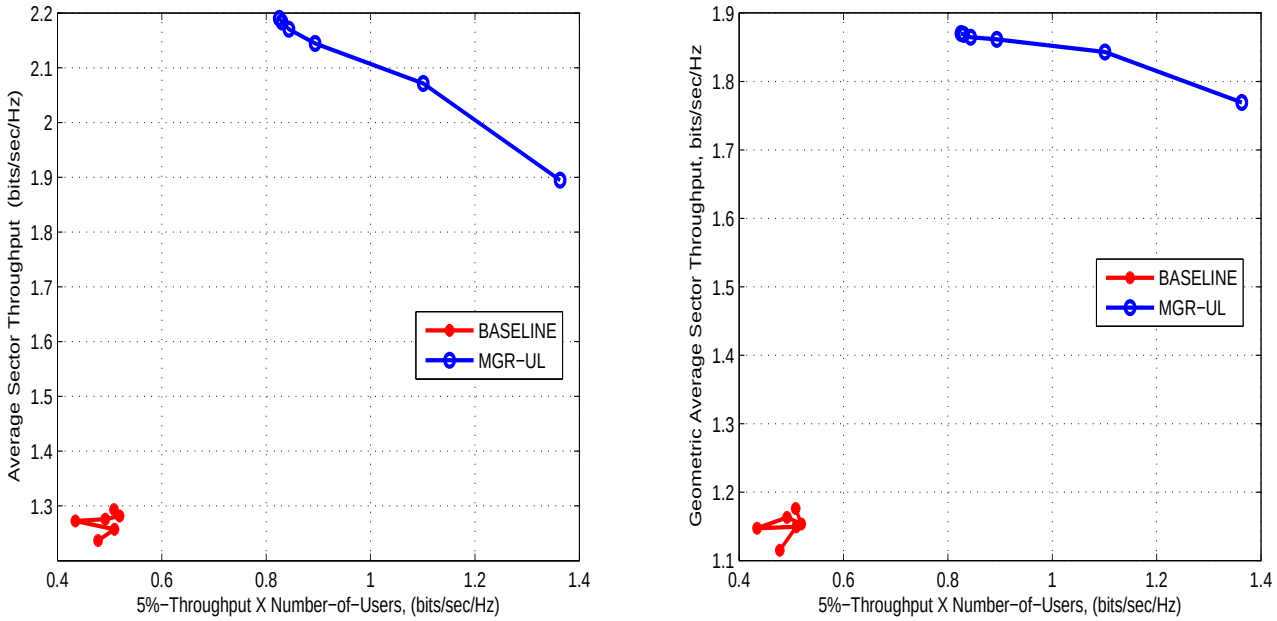

Figure 3: Average and geometric average sector throughputs Vs. Normalized 5-% edge throughput. No fast fading, 6 subbands

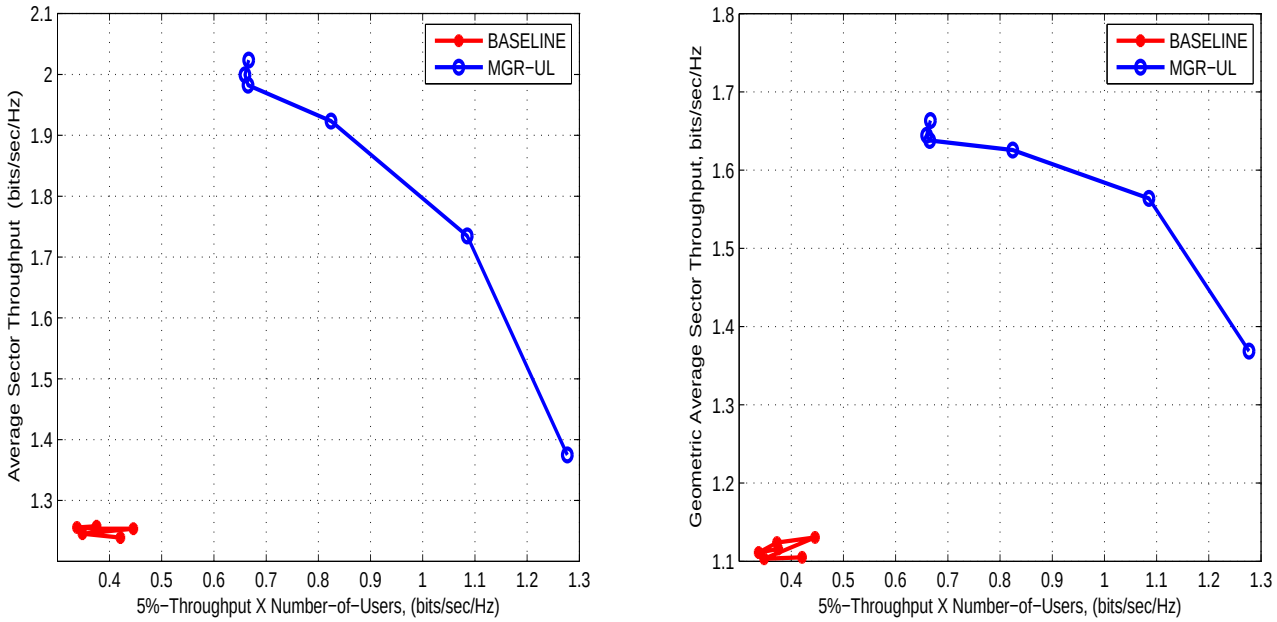

Figure 4: Average and geometric average sector throughputs Vs. Normalized 5-% edge throughput. No fast fading, 1 subband

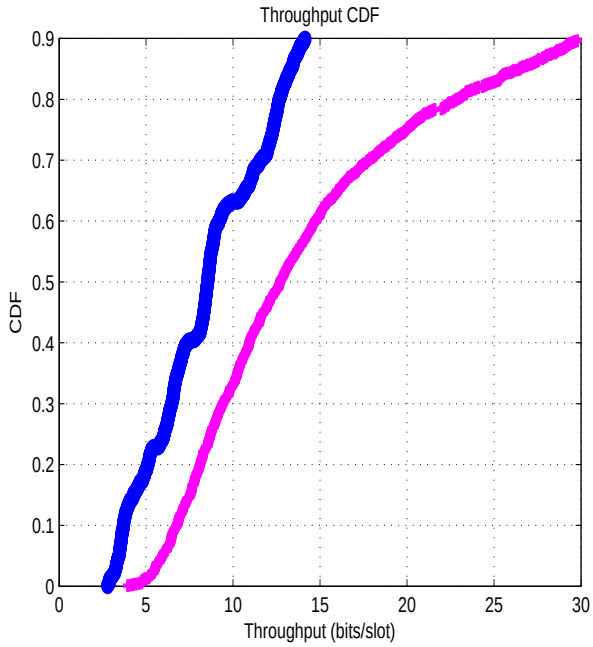

Figure 5: Cumulative distribution function of user throughputs. No fast fading, 6 subbands

6.3 Results and discussion

Figure 3 shows the performance of the MGR-UL algorithm relative to the Baseline algorithm for the case of no fast fading. (The x-axis is the normalized 5%-ile of the user throughput distribution across the system – this is a measure of "cell edge throughput". On the y-axis we show both the average sector throughput and geometric average sector throughput; the latter is defined as the geometric average of user throughputs across entire system, times the number of users per sector. The reason we are interested in geometric average sector throughput is because maximizing this metric is equivalent to maximizing the system utility chosen for our simulations, which is the sum of user log-throughputs.) Different points on the curves correspond to different values of minimum rate parameter R_{min}, which range from 0 to 10 bits/slot. From the figure, we observe that the MGR-UL algorithm provides substantial improvement in both the average sector throughput and the 5-percentile throughput. The sector throughput improvement, corresponding to $R_{\min} = 0$, is 69% while the highest edge throughput improvement is 176% with $R_{\min} = 10$.

Notice that for the Baseline scheme chaning R_{min} parameter has little effect on the system performance. In particular, increasing R_{min} practically does not increase cell edge user throughput. This is because Baseline, which by definition tries to equalize *all* user SINRs within a sector, has a tendency towards equalizing user throughputs as well; as a result, attempts to increase edge user throughputs within the framework of Baseline are ineffective. On the other hand, the interference-cost based power setting in MGR-UL allows "good" users (those close to serving base station) to achieve much higher SINRs. This not only improves overall spectral efficiency, but also allows to "free up" more or less time-frequency resources for the cell edge users, when necessary, thus allowing a trade-off between total and cell edge throughputs.

The performance improvement from MGR-UL comes because of two effects, which can be roughly

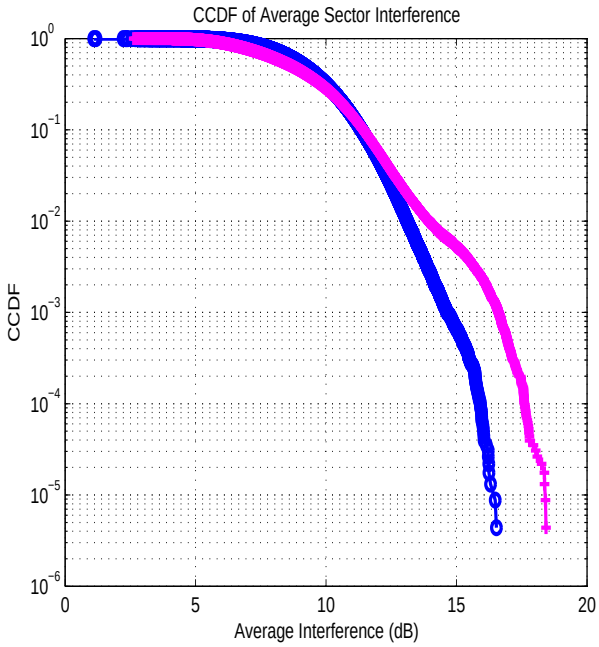

Figure 6: Distribution of average sector interference. No fast fading, 6 subbands

thought of as separate. The first is an optimized, cost based, setting of mobile transmit powers, which is applicable even in the single sub-band case (i.e., no FFR is used). The second effect, which works "on top of" the first, is an "appropriate" (also cost based) assignment of mobiles in different sectors to different sub-bands. To isolate the benefit of optimized power allocation alone from the benefit of jointly optimized power and sub-band allocation, we ran the MGR-UL algorithm for a single sub-band. The results are shown in Figure 4. Comparing the results of six sub-bands with that of the single sub-band for this specific scenario, we see that a significant part of the total benefit of MGR-UL may come from optimized power allocation alone. (We want to emphasize, however, that this observation does not necessary hold across all possible scenarios.)

Figure 5 shows the cumulative distribution function of the user throughputs for the MGR-UL and the Baseline algorithms for the six sub-band case. From the curves it is clear that the MGR-UL algorithm improves throughput performance for all users and is not simply trading off throughput of "good" users with that of cell edge users.

For the Baseline algorithm, one of the parameters is the average interference levels at the base station. This is typically quantified in OFDMA systems through Interference-over-thermal (IoT) values, which is the ratio of the average interference to thermal noise power. Enforcing a low value of IoT will curtail mobile transmit powers and result in reduced throughputs. In such a case, throughput can simply be increased by increasing the IoT threshold. Thus, a fair comparison of different algorithms should also include a comparison of the IoT values achieved by the algorithms. This is shown in Figure 6 as complementary cumulative distribution functions (CCDF) of the IoT values across the different sectors and over different slots. From the figure it is clear that the IoT values are comparable and thus a comparison of throughputs is fair.

Figure 7 shows the average interference across the six sub-bands for the three co-located sectors.

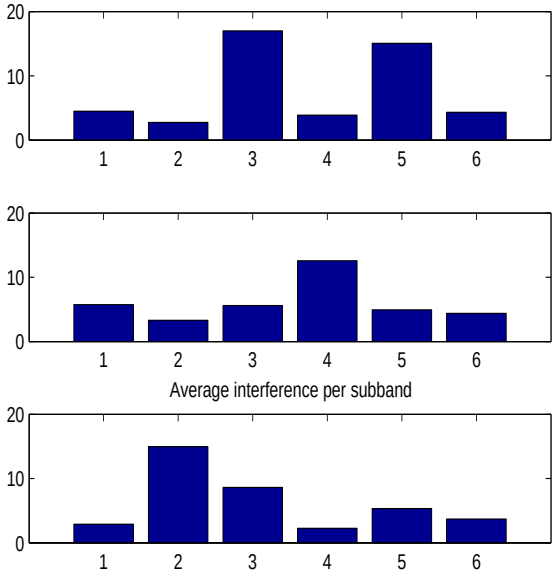

Figure 7: Average interference per subband in sectors 1-3. No fast fading, 6 subbands

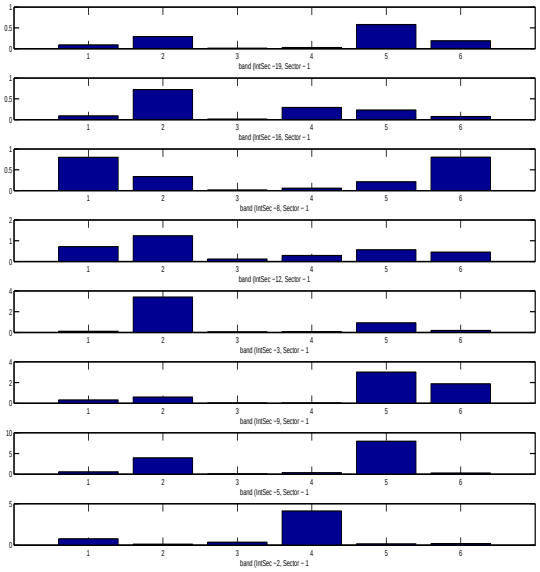

Figure 8: Average interference from sector 1 to neighbors, per subband. No fast fading, 6 subbands

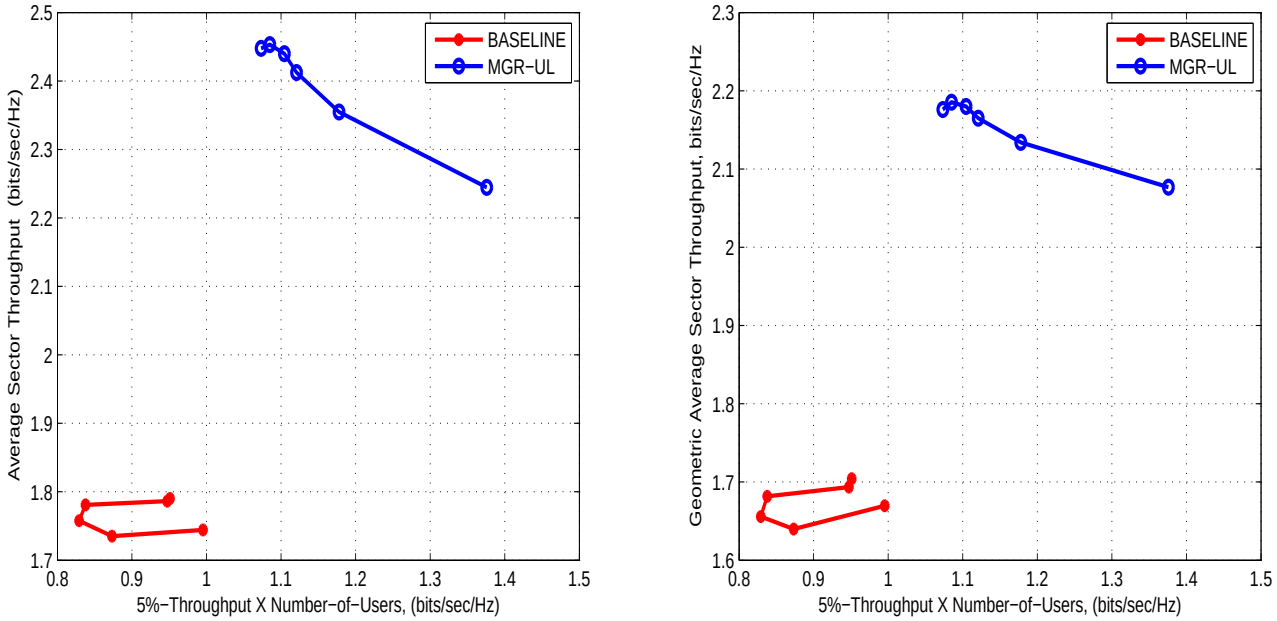

Figure 9: Average and geometric average sector throughputs Vs. Normalized 5-% edge throughput. Fast fading, 6 subbands

It is clear that the MGR-UL algorithm indeed results in a soft reuse pattern where some sub-bands are preferred in certain sectors. In these preferred sub-bands, these sectors generate a significant amount of out-of-cell interference to neighboring sectors. This is further confirmed in Figure 8 where the average interference caused by sector 1 to other neighboring sectors is shown for each sub-band.

While we have chosen to illustrate the basic results for the case of no fast fading, we show the average sector throughput and geometric sector throughput comparisons for the case with fast fading in Figure 9. Results show that MGR-UL can achieve either the average throughput improvement of 36% (with a small cell edge throughput improvement as well), or the edge throughput improvement of 38% (with still a 20% average throughput increase).

7 Conclusions and future work

We proposed a resource allocation scheme for the uplink of OFDMA system, whose underlying objective is to continuously maximize the overall system utility, which in turn depends on the achieved user throughputs. "In the process of doing that", the scheme dynamically creates an efficient FFR pattern, appropriate for each given spatial user traffic distribution. The scheme is semi-autonomous, with the inter-cell communication limited to infrequent exchange of interference costs. Our simulations show that the scheme achieve substantial performance improvements over a baseline scheme.

We observed that the performance gains are larger in the case of no fast fading – this is similar to the situation with downlink dynamic FFR schemes. Another observation is that, in some scenarios, a substantial part of the total gain can be achieved by using interference-cost based power assignment alone (without FFR) – when this happens (and can be detected) using such simplified scheme may

be attractive in practice.

The idea of using interfence costs (to neighbors) can potentially be used for the purposes other than dynamic FFR. For example, using constant, a priori fixed, interference costs amounts to a form of *static* FFR; in this case *no* inter-cell communication is required. (This may be an attractive option, if inter-cell exchange is undesirable or infeasible.) It might be of interest to compare the performance of such static scheme to that of other static schemes proposed in the literature.

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*, 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] 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.

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

[10] A. L. Stolyar, H. Viswanathan, "Self-organizing Dynamic Fractional Frequency Reuse for Best-Effort Traffic Through Distributed Inter-cell Coordination," in *Proceedings of INFOCOM'2009*, Rio-de-Janeiro, April 19-25, 2009.

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

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

[13] Video Share from AT&T, http://www.wireless.att.com/learn/messaging-internet/media-entertainment/attvideoshare.jsp