

# On Exploiting Diversity and Spatial Reuse in Relay-enabled Wireless Networks

Karthikeyan Sundaresan, and Sampath Rangarajan  
Broadband and Mobile Networking, NEC Laboratories America  
karthiks@nec-labs.com, sampath@nec-labs.com

## ABSTRACT

Relay-enabled wireless networks (eg. WIMAX 802.16j) represent an emerging trend for the incorporation of multi-hop networking solutions for last-mile broadband access in next generation wireless networks. The adoption of more sophisticated access technologies such as OFDM (orthogonal frequency division multiplexing) coupled with the relay-induced two-hop nature, provides two key benefits to these networks in the form of *diversity* and *spatial reuse* gains. However, leveraging these benefits calls for more sophisticated solutions, among which, user scheduling forms a key component.

We consider the specific problem of scheduling users with finite buffers on the multiple OFDM carriers (channels) over the two hops of the relay-enabled network. We propose scheduling algorithms that help leverage diversity and spatial reuse gains from these networks. We show that even the scheduling problem to exploit diversity gains alone is NP-hard and provide both theoretically and practically efficient polynomial-time algorithms with approximation guarantees. Building on the diversity solutions, we also propose an efficient polynomial-time scheduling algorithm for exploiting both spatial reuse as well as diversity. The proposed solutions are evaluated to highlight the relative significance of diversity and spatial reuse gains with respect to varying network conditions.

## Categories and Subject Descriptors

C.2.1 [Network Architecture and Design]: Wireless Communication

## General Terms

Algorithms, performance, theory

## 1. INTRODUCTION

The last decade has seen a significant amount of research in multi-hop wireless networks (MWNs) [1, 2]. While their completely decentralized nature has contributed to scalable solutions, they have also faced significant challenges in moving towards commercial adoption. However, with the next generation wireless net-

works moving towards smaller (micro, pico) cells for providing higher data rates, there is a revived interest in MWNs from the perspective of integrating them with infrastructure wireless networks. With a decrease in cell size, relays are now needed to provide extended coverage, resulting in a multi-hop network. A *two-hop relay-enabled* wireless network forms an important first step towards such a network model (Figure 1(a)). Here, the relay stations (RS) are connected to the wireless infrastructure (base station, BS) and provide improved coverage and capacity to several applications including serving mobile users (MS) in business hot-spots, office buildings, transportation vehicles, coverage holes, etc. The plethora of envisioned applications has also led to their adoption as the mandatory network model in IEEE 802.16j amendment to the WIMAX standard and in turn forms the context for our work.

Orthogonal frequency division multiplexing (OFDM) has become the popular choice for air interface technology in future local and wide area wireless networks. The entire spectrum is divided into multiple carriers (sub-channels), leading to several physical layer and scheduling benefits [3, 4]. The two-hop network model coupled with OFDM provides two key benefits, namely *diversity* and *spatial reuse* gains. Three kinds of diversity gains can be exploited through scheduling: (i) *multi-user diversity*: for a given sub-channel, different users experience different fading statistics, allowing us to pick a user with a larger gain; (ii) *channel diversity*: sub-channels experiencing high gain could vary from one user to another, allowing for multiple users to be assigned their best channels in tandem; and (iii) *cooperative diversity*: relays can exploit wireless broadcast advantage to cooperate and improve the SNR (signal-noise ratio) at the intended receiver. In addition to the diversity gain, the two-hop network model also provides room for *spatial reuse*, whereby simultaneous transmissions on the relay hop (BS-RS) and access hop (RS-MS) can be leveraged on the same channel as long as there is no mutual interference.

User and channel diversity gains, available in conventional one-hop cellular networks, have been effectively leveraged to improve system performance through several channel-dependent scheduling schemes [3, 4, 5]. However, they do not provide spatial reuse or cooperative diversity gains. MWNs on the other hand, provide spatial reuse. However, since diversity gains require channel state feedback from RS and MS and must be exploited at fine time scales (order of frames), they cannot be effectively leveraged in a large multi-hop setting. Relay-enabled networks with a two-hop structure, provide a unique middle-ground between these two networks, providing us access to a multitude of diversity and spatial reuse gains. While this provides potential for significant performance improvement, it also calls for more sophisticated, tailored scheduling solutions that take into account the two-hop nature of the system. In this context, we focus on the specific problem of scheduling users

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MobiHoc'08, May 26–30, 2008, Hong Kong SAR, China.  
Copyright 2008 ACM 978-1-60558-073-9/08/05...\$5.00.

with finite buffers on multiple OFDM sub-channels over the two hops of the network, while efficiently exploiting the available diversity and spatial reuse gains. Leveraging just the diversity gains over two hops is an important problem in itself, whose hardness we first establish. Then we make the following contributions.

- We provide two theoretically efficient polynomial-time scheduling algorithms for exploiting diversity gains with approximation guarantees of  $\left(1 - \sqrt{\frac{c'K}{N} \cdot \log(2N)}\right)$  and  $\left(\left(1 - \frac{1}{e}\right)^2 - \epsilon\right)$  in the order of increasing time-complexity.  $K$  and  $N$  are the number of users and sub-channels respectively, and  $c' > 1, \epsilon > 0$  are constants. We also provide a practically efficient adaptive scheduling algorithm with good average case performance.
- Building on the diversity solutions, we also propose an efficient polynomial-time scheduling algorithm for exploiting both spatial reuse and diversity.
- The proposed solutions are evaluated to highlight the relative significance of diversity and spatial reuse gains with respect to varying network conditions.

The rest of the paper is organized as follows. The system description is presented in Section 2. Sections 3 and 4 present the proposed scheduling algorithms for exploiting diversity and spatial reuse respectively. Performance evaluation is presented in Section 5, followed by some discussions in Section 6. Finally, concluding remarks are presented in Section 7.

## 2. SYSTEM DESCRIPTION

### 2.1 Related Work

Several works [6, 7] have investigated the potential of relay-enabled wireless networks to provide improved coverage and capacity. These networks have gained attention both in the standards community (IEEE 802.16j) and in the industry [8]. Scheduling, being a key component in these networks, has received higher emphasis [9, 7, 10]. However, most of these works [9, 7] focus on TDMA variants where the scheduling decision reduces mainly to deciding whether to employ a relay or not and for which particular user. They focus on link level performance and do not exploit spatial reuse and diversity across the relay and access hops that is available at a network level. Further, they [9, 7, 10] do not consider a multiple channel OFDM network (channel diversity), which complicates scheduling decisions with the possibility of multiple users operating in parallel. The works on OFDM scheduling in conventional cellular systems [3, 4] cannot be directly applied to two-hop relay networks, where the network structure is different and spatial reuse and diversity across hops forms an important component.

There have been some works [11, 12] that have looked at multiple channels in the presence of relays, where reassignment of channels at the second hop is considered to exploit diversity better. However, the channel assignment model considered is simplistic and also spatial reuse is not exploited. On the other hand, works in MWNs [1, 2] leverage spatial reuse, but they do not focus on OFDM and associated diversity gains due to the large scale. Also, none of the above works take into account the finite (non-backlogged) user buffers at the BS, which makes the problem NP-hard. This was recently considered in [5] in the single hop context. We consider leveraging both diversity and spatial reuse gains in the presence of finite user buffers in a more difficult two-hop setting.

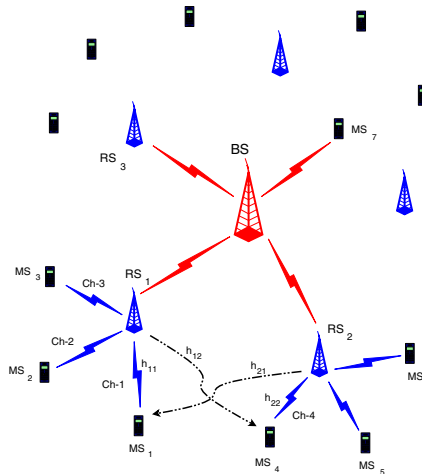


Figure 1: Network Model

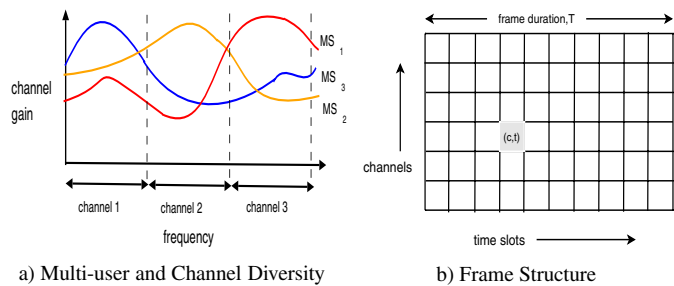


Figure 2: Potential Gains and Frame Structure.

### 2.2 Network Model

We consider a downlink OFDMA-based, relay-enabled, two-hop wireless network as shown in Figure 1. A set of  $K$  mobile stations (MS) are uniformly located within an extended cell radius. A small set of  $R$  relay stations (RS) are added to the mid-way belt of the network. MS that are closer to the BS directly communicate with it. However, MS farther from the BS connect with the RS that is closest to them. The one-hop links between BS and RS are referred to as *relay links*, RS and MS as *access links*, and BS and MS as *direct links*. The BS, RS and MS are allowed to operate on multiple channels from a set of  $N$  total OFDM sub-channels. Data flows are considered and assumed to originate in the Internet and destined towards the MS. Let  $P$  denote the maximum power used by the BS for its transmission, which is split equally across all sub-channels and no power adaptation across channels is assumed, given the marginal gains resulting from it [13]. Note that a sub-channel could correspond to a single carrier or a bunch of contiguous carriers as in practical systems. RS (MS) are assumed to provide feedback of their relay (access) channel rates to BS (RS). All stations are assumed to be half-duplex. Hence, an RS can be active on only its relay or access link in any slot but not both.

### 2.3 Potential Gains

Relay networks provide two key benefits, namely diversity (link-level) and spatial reuse (network-level) gains. Three forms of diversity gains are possible. Consider the frequency response of three channels for three MS in Figure 2(a). Multipath fading and user mobility result in independent fading across users for a given chan-

nel, contributing to *multi-user diversity*. Further, the presence of multiple channels and the corresponding frequency selective fading results in different channels experiencing different gains for a given MS, contributing to *channel diversity*. These gains make it possible to schedule multiple users in tandem, while providing good quality channels to many of them (eg. channels 3, 2 and 1 to MS 1, 2 and 3 respectively).

Consider data  $x_1$  and  $x_2$  transmitted by  $BS$  on channels 1 and 2 towards  $RS_1$  and  $RS_2$ , and destined for  $MS_1$  and  $MS_4$  respectively in Figure 1. Due to wireless broadcast advantage,  $RS_1$  and  $RS_2$  receive both data from the BS. Now, in addition to  $RS_1$  ( $RS_2$ ) forwarding  $x_1$  ( $x_2$ ) on channel, say 1 (2) to  $MS_1$  ( $MS_4$ ), it can also cooperate to transmit a coded version of the unintended data  $x_2$  ( $x_1$ ) on the other channel 2 (1) to increase the SNR at  $MS_4$  ( $MS_1$ ), thereby resulting in *cooperative diversity* gain. Since our focus is only to enable our scheduler leverage cooperative gains if made available, we consider a simple transmit beamforming scheme for relay cooperation, and assume only the two neighboring, adjacent RS participate in cooperation with the desired RS. The access rates fed back from MS are assumed to incorporate the effect of cooperation and fading.

The spatial separation of RS allows relay and access links to operate in tandem, *spatially reusing* the same set of channels across hops without causing mutual interference (eg.  $BS$ - $RS_3$  and  $RS_2$ - $MS_5$  operating in parallel). We do not consider reuse of channels within the access hop, since it does not lead to benefits unless the access hop becomes the network bottleneck, which is usually not the case. Further, it comes at the cost of spatial reuse across hops, which is a more important feature to be leveraged. However, we do outline later how it can be incorporated into our solutions.

## 2.4 Scheduling Model

We consider a synchronized, time-slotted system similar to a WiMAX relay, with  $BS$  and  $RS$  transmitting data in frames. Every frame consists of several time slots and has to be populated with user assignments across both time slots and channels as in Figure 2(b). It is sufficient to consider the problem with one time slot per frame since channels in other time slots can be considered as additional channels available to the considered time slot. The  $RS$  assignments to relay channels for the current frame and the  $MS$  assignments to access channels for the next frame are indicated to  $RS$  through a MAP that follows the preamble in the frame (eg. 802.16j). In every slot of a frame, a set of  $RS$  and/or  $MS$  on the relay and access hops respectively are activated based on the assignments provided by the  $BS$ . For ease of exposition, we present our discussions with respect to only relay and access links. Direct links can be easily incorporated into the solutions by considering them as relay links without affecting performance guarantees.

Thus, the scheduling problem is now to find a feasible assignment of relay channels to  $RS$  on the relay hop, along with a feasible assignment of access channels to  $MS$  on the access hop such that the desired objective of the system is maximized while conforming to all constraints. The constraints include the finite user buffers at  $BS$ , assignment of a relay (access) channel to at most one  $RS$  ( $MS$ ), and half-duplex constraint of all stations. The objective of our scheduling algorithms is to maximize the end-to-end system throughput subject to a desired fairness model. We consider the proportional fairness model, given its ability to strike a good balance between utilization and fairness [14]. The considered system can be shown to converge to the optimum if the scheduler's decisions at each epoch (interval) are made to maximize the aggregate *marginal* (incremental) utility,  $S_{\max} = \arg \max_S \left\{ \sum_{j \in S} \Delta U_j \right\}$ .  $\Delta U_j$  denotes the marginal flow (two-hop) utility received by user

$j$  in a feasible schedule  $S$  and is given by  $\frac{\beta_j r_j^{eff}}{\bar{r}_j}$  for proportional fairness, where  $\beta_j$  captures the priority weight of user's QoS class and  $\bar{r}_j$  its average throughput. The average throughput of a user is kept track at the  $BS$  using weighted averaging techniques similar to those considered in single channel systems.

$r_j^{eff}$  corresponds to the user's *two-hop flow rate*, which in turn is determined by the instantaneous *effective* rate on the relay and access hops combined. Let  $r_j^{rel}$  and  $r_j^{acc}$  be the net bit-rates obtained for a user (flow)  $j$  on the relay and access hops respectively, and  $B_j$  be the available data in user  $j$ 's buffer at the  $BS$ . No per-user buffer is assumed at the simplistic  $RS$ . Then  $r_j^{eff} = \frac{1}{2} \cdot \min \left\{ r_j^{rel}, r_j^{acc}, \frac{B_j}{T} \right\}$ , where  $T$  is the frame duration. In the case of backlogged user buffers, the effective rate can be given by,  $r_j^{eff} = \frac{\min \{ r_j^{rel}, r_j^{acc} \}}{2}$ . The net rates obtained on the relay and access hops are in turn determined by the set of relay ( $A_j^{rel}$ ) and access ( $A_j^{acc}$ ) channels assigned to the user, namely  $r_j^{rel} = \sum_{n \in A_j^{rel}} r_{n,j}^{rel}$  and  $r_j^{acc} = \sum_{m \in A_j^{acc}} r_{m,j}^{acc}$ . Irrespective of the metric used, it must be noted that the effective rate cannot be decoupled into independent components of the relay and access part, thereby inherently necessitating joint two-hop decisions.

## 3. DIVERSITY SCHEDULING

We first consider exploitation of diversity gains alone, which is an important problem in itself in addition to constituting an important step towards our joint diversity and spatial reuse solution. Since simultaneous transmissions (spatial reuse) are not leveraged on the same channel, the relay and the access hops are scheduled sequentially in two slots, satisfying the half-duplex constraint of  $RS$ . We allow the data that is meant for a  $MS$ , to be sent using different sets of channels on the relay and access hops. This allows for diversity gains *across* hops in addition to those available within hops. The problem is now essentially to find a two-slot schedule, namely an assignment of  $N$  channels to  $RS$  and  $MS$  on the relay and access hops respectively, such that the aggregate marginal utility of the system ( $K$  users/flows) is maximized at every scheduling epoch. We start with a relaxed set of constraints, where the problem can be solved optimally in polynomial time. Then we consider the comprehensive model and propose approximation solutions.

### 3.1 Pair-wise Channel Assignment (PCA)

Relay and access channels are assigned to a user in pairs  $(n, m)$  with the contribution of individual pairs to a user assumed to be independent. Thus, while a user can be assigned several channel pairs, the total number of channels assigned on the relay and access hops is equal. Further, backlogged user buffers is assumed. The marginal utility of a user (flow)  $j$  on a relay-access channel pair  $(n, m)$  across two hops (slots) is given by  $\frac{\beta_j r_{n,m,j}^{eff}}{\bar{r}_j}$ , where  $r_{n,m,j}^{eff} = \frac{\min \{ r_{n,j}^{rel}, r_{m,j}^{acc} \}}{2}$ . The pair-wise restriction reduces channel diversity gains, with inefficiencies also introduced by the backlogged buffer assumption. However, such a simple model is extremely attractive for practical real-time implementation, and hence it is important to leverage it if it can provide good performance even under certain restricted conditions. Under this model, the scheduling problem can be stated as,

$$S_{\max}(t) = \arg \max_S \left\{ \sum_{j \in S} \frac{\beta_j}{\bar{r}_j(t)} \sum_{n=1}^N \sum_{m=1}^N r_{n,m,j}^{eff}(t) I_{n,m,j}(t) \right\}$$

$$\sum_{j=1}^K \sum_{n=1}^N I_{n,m,j}(t) \leq 1, \quad \forall m; \quad \sum_{j=1}^K \sum_{m=1}^N I_{n,m,j}(t) \leq 1, \quad \forall n$$

where  $I_{n,m,j}(t) \in \{0, 1\}$ , is a binary function capturing the assignment of (relay,access) channel pair  $(n, m)$  to user  $j$  in epoch  $t$ . The two constraints indicate that every channel on either hop can at most be assigned to a single user. The net effective rate assigned to a user is  $r_j^{eff} = \sum_{(n,m)} r_{n,m,j}^{eff} I_{n,m,j}(t)$ . The above problem is solved optimally by solving an equivalent maximum utility bipartite matching problem using the following algorithm DIV1.

- 
- Construct a bipartite graph:  $G = (V_1 \times V_2, E)$ , where the vertices in  $V_1$  and  $V_2$  correspond to the set of sub-channels on the relay and access hops with  $|V_1| = |V_2| = N$ . The edge set  $E$  corresponds to  $N^2$  edges connecting all possible pairs of channels in the two sets.
  - The weight of every edge carries two attributes,  $(w_{ik}, u_{ik})$ , where  $u_{ik}$  is the user providing the maximum marginal utility ( $w_{ik}$ ) for the channel pair  $(i, k)$ , namely  $w_{ik} = \max_j \left\{ \frac{\beta_j r_{i,k,j}^{eff}}{\bar{r}_j} \right\}$  and  $u_{ik} = \arg \max_j \left\{ \frac{\beta_j r_{i,k,j}^{eff}}{\bar{r}_j} \right\}$ . Using marginal utilities as the weights takes into account the average throughput of users and hence fairness.
  - Run a maximum weight bipartite matching algorithm on  $G$  to obtain the set of  $N$  channel pair assignments on relay and access links, providing the maximum marginal utility.

---

Further, the second attribute of the edges present in the maximum matching provides the set of RS and associated MS to be scheduled over two consecutive slots: relay links followed by the access links. Several good polynomial-time algorithms exist for solving the bipartite matching problem, of which the Hungarian algorithm [15] is a popular choice.

### 3.2 Flexible Channel Assignment (FCA)

The restriction on number of channels assigned to a user on relay and access hops being equal is now removed, thereby allowing for maximum channel diversity gains. Further, finite data in users' buffers at the BS is taken into account in scheduling. Unlike in the previous model, where the net effective rate of a user was determined by the sum of the rates on the assigned channel pairs ( $r_j^{eff} = \sum_{(n,m) \in A_j} r_{n,m,j}^{eff}$ ), here the effective rate is calculated based on the net flow that can be sent to the user, while accounting for finite

user buffer at BS ( $r_j^{eff} = \frac{\min \left\{ \sum_{n \in A_j^{rel}} r_{n,j}^{rel}, \sum_{m \in A_j^{acc}} r_{m,j}^{acc}, \frac{B_j}{T} \right\}}{2}$ ). The incorporation of either flexible channel assignment (across hops) or finite user buffer is sufficient to make the problem hard.

**THEOREM 1.** *The diversity problem under the flexible assignment model is NP-hard.*

**PROOF.** (Sketch) The problem can be reduced from two-dependent multiple knapsack problems (MKP), where each set of knapsacks (bins) belongs to a different type (hop). When the profit of an item (channel) depends on only the bin (user) it is assigned to and not the type (hop) of the bin, the problem reduces to a single instance of MKP, which by itself is known to be NP-hard and even hard to approximate [16], indicating that the type-dependent profit (hop-dependent channel rate) is a much harder version.  $\square$

Given the hardness of the problem, we now provide two solutions based on LP relaxation and rounding, with provable performance bounds in the order of increasing performance, but also increasing complexity.

#### 3.2.1 Algorithm DIV2

Let  $x_{ij}$  and  $y_{kj}$  be integer variables indicating the assignment of relay channel  $i$  and access channel  $k$  to user  $j$  respectively;  $w_{ij}^r$  and  $w_{kj}^a$  denote the one-hop marginal utility (normalized to a large value  $W_{max}$ ) obtained by user  $j$  on being assigned relay channel  $i$  ( $\beta_j r_{i,j}^{rel} / \bar{r}_j$ ) and access channel  $k$  ( $\beta_j r_{k,j}^{acc} / \bar{r}_j$ ) respectively;  $a_j$  denotes the aggregate marginal utility (flow) achieved by user  $j$ , with  $B_j$  and  $\bar{r}_j$  representing the data (buffer) availability and average throughput of user  $j$ , and  $T$  being the frame duration. FCA can now be modeled using the following mixed integer program.

$$\begin{aligned} & \text{Maximize } \sum_{j \in U} a_j & (1) \\ \text{subject to } & \sum_{j \in U} x_{ij} \leq 1, \quad \forall i \in C_r; \quad \sum_{j \in U} y_{kj} \leq 1, \quad \forall k \in C_a \\ & \sum_{i \in C_r} w_{ij}^r x_{ij} = a_j; \quad \sum_{k \in C_a} w_{kj}^a y_{kj} = a_j, \quad \forall j \in U \\ & a_j \leq \frac{\beta_j B_j}{\bar{r}_j T}, \quad \forall j \in U \\ & x_{ij}, y_{kj} \in \{0, 1\}; \quad a_j \geq 0; \quad w_{ij}^r, w_{kj}^a \in [0, 1], \quad \forall i, j, k \end{aligned}$$

The first set of constraints allows for at most one user assignment to any channel on relay and access hops. The second set represents flow (marginal utility) conservation for each user at its associated RS, while the last captures the finite buffer limit for each user. The algorithm, DIV2 is as follows.

---

**Step 1:** Formulate FCA using the MIP above. Solve its LP relaxation with  $x_{ij}$  and  $y_{kj}$  being relaxed to  $[0, 1]$ . Let the LP solutions be  $x_{ij}^*$ ,  $y_{kj}^*$  and  $a_j^*$ .

**Step 2:** Adopt the following procedure to round the fractional solutions,  $x_{ij}^*$  and  $y_{kj}^*$  to integral values,  $\hat{x}_{ij}$  and  $\hat{y}_{kj}$ .

- For every relay channel  $i$ , round  $x_{ij}$  to 1 ( $\hat{x}_{ij}$ ) with probability  $x_{ij}^*$ . If  $j^*$  is the user to whom relay channel  $i$  is assigned, then  $\hat{x}_{ij} = 0, \forall j \neq j^*$ .
- Update the final constraint as,  $a_j \leq \min \left\{ \frac{\beta_j B_j}{\bar{r}_j T}, \frac{\sum_{i \in C_r} w_{ij}^r \hat{x}_{ij}}{1 - \delta} \right\} \forall j$ , where  $\delta$  is a constant (derived below). Run the LP on only the  $y_{kj}$  variables. Let  $\bar{y}_{kj}$  and  $\bar{a}_j$  be the solutions of this new LP.
- For every access channel  $k$ , round  $y_{kj}$  to 1 ( $\hat{y}_{kj}$ ) with probability  $\bar{y}_{kj}$ . If  $j^*$  is the user to whom access channel  $k$  is assigned, then  $\hat{y}_{kj} = 0, \forall j \neq j^*$ .

---

The above rounding procedure ensures that any channel is assigned to at most one user. However, it is possible that the flow conservation at a user is violated. Further, the flow assigned to a user may also violate its buffer limit,  $\sum_{i \in C_r} w_{ij}^r \hat{x}_{ij} > \frac{\beta_j B_j}{\bar{r}_j T}$  or  $\sum_{k \in C_a} w_{kj}^a \hat{y}_{kj} > \frac{\beta_j B_j}{\bar{r}_j T}$ , using more data than what is actually available in the buffer. In such cases, the resulting flow of the user in the rounded solution is given by  $\min \left\{ \sum_{k \in C_a} w_{kj}^a \hat{y}_{kj}, \sum_{i \in C_r} w_{ij}^r \hat{x}_{ij}, a_j^* \right\}$ . We now capture the performance of the algorithm.

**THEOREM 2.** *Algorithm DIV2 provides an approximation guarantee of at least  $1 - \sqrt{\frac{e'K}{N}} \log(2N)$  with high probability.*

**PROOF.** We use the Chernoff-Hoeffding bound to bound the probability of buffer violations.

[Hoeffding-Chernoff Bound]: For  $v_i \in [0, 1]$  independent random variables (R.V's), let  $S = \sum_i v_i$ ,  $\mu = E[\sum_i v_i]$ , then

$$\Pr[S \leq (1 - \delta)\mu] \leq e^{-\frac{\delta^2 \mu}{2}} \quad (2)$$

Now, let  $\Phi_{x,j} = \sum_i w_{ij}^r \hat{x}_{ij}$  and  $\Phi_{y,j} = \sum_k w_{kj}^a \hat{y}_{kj}$  be the random variables denoting the flow (aggregate marginal utility) assigned to user  $j$  on the relay and access hops respectively in the rounded solution. Both  $\Phi_{x,j}$  and  $\Phi_{y,j}$  are sums of R.V's  $\in [0, 1]$ . With  $E[\Phi_{x,j}] = a_j^*$ , the probability that the relay flow at a user is atleast  $1 - \delta$  of the optimal can be obtained from equation 2 as,

$$\Pr[\Phi_{x,j} \geq (1 - \delta)a_j^*] \geq 1 - e^{-\frac{\delta^2 a_j^*}{2}} \quad (3)$$

Now, the flow on the access hop is obtained based on the flow in the rounded solution from the relay hop. Without loss of generality, assume  $\frac{\sum_{i \in C_r} w_{ij}^r \hat{x}_{ij}}{1 - \delta} \leq \frac{\beta_j B_j}{r_j T}$ . Given a relay flow of  $\sum_{i \in C_r} w_{ij}^r \hat{x}_{ij} \geq (1 - \delta)a_j^*$ , the access flow is bounded by  $\frac{\sum_{i \in C_r} w_{ij}^r \hat{x}_{ij}}{1 - \delta}$ , where  $\frac{\sum_{i \in C_r} w_{ij}^r \hat{x}_{ij}}{1 - \delta} \geq a_j^*$ . Further, an access flow of  $a_j^*$  is achievable (from first LP), resulting in  $\bar{a}_j \geq a_j^*$ . Since  $E[\Phi_{y,j} | \Phi_{x,j}] = \bar{a}_j \geq a_j^*$ , it can be shown that,

$$\Pr[\Phi_{y,j} \geq (1 - \delta)a_j^* | \Phi_{x,j}] \geq 1 - e^{-\frac{\delta^2 a_j^*}{2}} \quad (4)$$

Now, to ensure a net flow of atleast  $(1 - \delta)a_j^*$  at every user with high probability, we need the dependent flow on both the hops to be atleast  $(1 - \delta)a_j^*$  with probability atleast  $1 - \frac{1}{N}$ . From equations 3 and 4, we now have,

$$\Pr[\Phi_{x,j}, \Phi_{y,j} \geq (1 - \delta)a_j^*] \geq \left\{ 1 - e^{-\frac{\delta^2 a_j^*}{2}} \right\}^2 = \left( 1 - \frac{1}{N} \right)$$

This results in a per-user loss factor of  $\delta = \frac{\gamma}{\sqrt{a_j^*}}$ , where  $\gamma = \sqrt{2 \cdot \log\left(\frac{\sqrt{N}}{\sqrt{N} - \sqrt{N-1}}\right)} < \sqrt{2 \log(2N)}$ . If  $OPT$  and  $\overline{OPT}$  are the objective values of LP relaxation and the rounded integral solution, then the approximation factor  $A$  is given by,

$$A = \frac{\overline{OPT}}{OPT} = \frac{\sum_{j \in U} (1 - \delta)a_j^*}{\sum_{j \in U} a_j^*} = 1 - \gamma \frac{\sum_{j \in U} \sqrt{a_j^*}}{\sum_{j \in U} a_j^*}$$

Since  $\frac{\sum_{j \in U} \sqrt{a_j^*}}{\sum_{j \in U} a_j^*}$  is maximum when  $a_j^* = a^* = \frac{\sum_{j \in U} a_j^*}{K} = \frac{OPT}{K}$ , now substituting for  $\gamma$ , we have  $A \geq 1 - \sqrt{\frac{2K}{OPT} \log(2N)}$ .

We know  $OPT \leq N$ . If  $c = \frac{\max_{i,k,j} \{w_{ij}^r, w_{kj}^a\}}{\min_{i,k,j} \{w_{ij}^r, w_{kj}^a\}} \geq 1$ , then  $OPT \geq \frac{N}{c}$ , resulting in  $A \geq 1 - \sqrt{\frac{2cK}{N} \log(2N)}$  whp.  $\square$

Thus, the approximation guarantee becomes better when there is a larger magnitude of available channels to users in the system, i.e.  $\frac{N}{\log(2N)} > K$ . Since the number of sub-channels in a practical OFDM system are much more than the number of users, DIV2 can be expected to provide good performance. However, when  $K$  is larger or comparable to  $N$ , the worst case performance can be arbitrarily bad. But the average-case performance can still be expected to be good due to the dependent-rounding employed across the two hops. To address this parameter-dependent performance, we present an alternate formulation with a *constant approximation guarantee*, albeit at the cost of increased complexity.

### 3.2.2 Algorithm DIV3

We now present an alternate formulation, extending the formulation presented in [16] for MKP (applicable to one-hop assignment) to our two-hop assignment problem. Let  $S_j$  for  $j \in U$  be the set of all feasible assignment of channels on relay and access hops for user  $j$ . We have  $s = \{s_r, s_a\}$  for  $s \in S_j$ , consisting of a pair of subsets, indicating the feasible assignment (subset) of channels on the relay ( $s_r$ ) and access ( $s_a$ ) hops respectively for a user. A subset pair  $s$  is said to be feasible for a user, if both the associated subsets ( $s_r, s_a$ ) individually satisfy the buffer constraint at the user. If  $w_j^s$  is the weight (marginal utility) of subset pair  $s$  assigned to user  $j$ , then we have  $w_j^s = \min\{\sum_{i \in s_r} w_{ij}^r, \sum_{k \in s_a} w_{kj}^a\}$ . We need to consider all possible subsets of channels on each hop since any subset can be converted to a feasible (buffer constraint satisfying) subset by adjustment of its weights as follows. Let  $w_j^s > \frac{\beta_j B_j}{r_j T}$  and let the relay hop subset  $s_r$  be the one providing the minimum of the marginal utilities. Now, update the weights as  $\hat{w}_{ij}^{s_r} = \frac{w_{ij}^r}{\sum_{i \in s_r} w_{ij}^r} \left( \frac{\beta_j B_j}{r_j T} \right)$  for  $i \in s_r$  and  $\hat{w}_{kj}^{s_a} = \frac{w_{kj}^a}{\sum_{k \in s_a} w_{kj}^a} \sum_{i \in s_r} \hat{w}_{ij}^{s_r}$  for  $k \in s_a$ . Then we have a feasible subset pair  $s = \{s_r, s_a\}$ , with  $\hat{w}_j^s = \sum_{i \in s_r} \hat{w}_{ij}^{s_r} = \sum_{k \in s_a} \hat{w}_{kj}^{s_a} = \frac{\beta_j B_j}{r_j T}$ . Further, let  $Z_j^s$  be the indicator variable that indicates the assignment of subset pair  $s$  to user  $j$ . The objective is to find an assignment of subset pairs to users such that the aggregate marginal utility of the system is maximized. The formulation is as follows.

$$\begin{aligned} & \text{Maximize} && \sum_{j \in U, s \in S_j} \hat{w}_j^s Z_j^s && (5) \\ & \text{subject to} && \sum_{j \in U, s \in S_j: c \in s} Z_j^s \leq 1, && \forall c \in [1, 2N] \\ & && \sum_{s \in S_j} Z_j^s = 1, && \forall j \in U \\ & && Z_j^s \in \{0, 1\}, && \forall j \in U, \forall s \in S_j \end{aligned}$$

The first constraint ensures that each channel on the relay ( $[1, N]$ ) and access ( $[N + 1, 2N]$ ) hops is assigned to atmost one of the subset pairs over all users, while the second ensures that each user is assigned only one subset. Non-scheduled users are assigned a null subset pair. The buffer constraints are implicitly incorporated in the weights. We now have the following algorithm, DIV3.

**Step 1:** Convert the IP formulation by relaxing variables  $Z_j^s \in [0, 1]$ . Solve the resulting LP relaxation. Let the fractional solution be  $Z_j^{s*}$ . Each  $Z_j^{s*}$  can be equivalently represented by two variables corresponding to the two subsets (relay and access hops) in  $s^*$ , namely  $Z_j^{s*} = \{X_j^{s_r^*}, Y_j^{s_a^*}\}$  with  $X_j^{s_r^*} = Y_j^{s_a^*} = Z_j^{s*}$  and optimal user flow  $\hat{w}_j^* = \sum_{s \in S_j^*} \hat{w}_j^s Z_j^{s*}$ .

**Step 2:** Use the following rounding procedure to round the fractional values to integral values  $\hat{X}_j^{s_r}$  and  $\hat{Y}_j^{s_a}$ .

- For every user  $j$ , round relay and access subset assignment variables  $X_j^{s_r}, Y_j^{s_a}$  to  $1 (\hat{X}_j^{s_r}, \hat{Y}_j^{s_a})$  simultaneously with probability  $Z_j^{s*}$ . Let the chosen subset pair be  $\hat{s} = \{\hat{s}_r, \hat{s}_a\}$ .  $\forall s_r \neq \hat{s}_r$ , set  $\hat{X}_j^{s_r} = 0$ . Similarly,  $\forall s_a \neq \hat{s}_a$ , set  $\hat{Y}_j^{s_a} = 0$ . Adjust the hop flows to the bottleneck flow s.t.  $\hat{w}_j^{s_r} = \hat{w}_j^{s_a} = \hat{w}_j^{\hat{s}}$  by updating weights as,  $\hat{w}_{ij}^{\hat{s}_r} \leftarrow \hat{w}_{ij}^{s_r} \left( \frac{\hat{w}_j^{\hat{s}}}{\hat{w}_j^{s_r}} \right)$ ,  $\forall i \in \hat{s}_r$  and  $\hat{w}_{kj}^{\hat{s}_a} \leftarrow \hat{w}_{kj}^{s_a} \left( \frac{\hat{w}_j^{\hat{s}}}{\hat{w}_j^{s_a}} \right)$ ,  $\forall k \in \hat{s}_a$ .
- After a (relay,access) subset pair is chosen for every user, it

is possible that a relay channel  $i$  ends up being assigned to multiple selected users (subsets), in which case the channel is assigned to the user ( $j^*$ ) with the largest weight,  $j^* = \arg \max_{j: \hat{X}_j^{\hat{s}_r} = 1, i \in \hat{s}_r} \hat{w}_{ij}^{\hat{s}_r}$ . Note that removing a channel from an assigned subset to a user does not violate buffer feasibility constraint. After channel feasibility is restored, for a given user, the updated (potentially reduced) relay subset is denoted by  $\bar{s}_r$  with weight  $\hat{w}_{ij}^{\bar{s}_r}$ .

- Based on relay flow, find a new set of access weights for each user as,  $\bar{w}_{kj}^{\hat{s}_a} = \hat{w}_{kj}^{\hat{s}_a} \left( \frac{\hat{w}_j^{\hat{s}_r}}{\hat{w}_j^{\bar{s}_r}} \right)$ ,  $\forall k \in \hat{s}_a, j \in U$ , where  $\hat{w}_j^{\bar{s}_r} = \sum_{i \in \bar{s}_r} \hat{w}_{ij}^{\bar{s}_r} \leq \hat{w}_j^{\hat{s}_r}$ . Hence,  $\bar{w}_{kj}^{\hat{s}_a} \leq \hat{w}_{kj}^{\hat{s}_a}$ , thereby also ensuring buffer feasibility.
- If access channel  $k$  is assigned to multiple users, use the new set of weights to break ties,  $j^* = \arg \max_{j: \hat{Y}_j^{\hat{s}_a} = 1, k \in \hat{s}_a} \bar{w}_{kj}^{\hat{s}_a}$ . Let  $\bar{s}_a$  be the updated access subset for a given user. The net flow for user  $j$  is  $\min\{\sum_{i \in \bar{s}_r} \hat{w}_{ij}^{\bar{s}_r}, \sum_{k \in \bar{s}_a} \bar{w}_{kj}^{\hat{s}_a}\}$ .

---

**THEOREM 3.** *The expected value of the rounded solution is atleast  $(1 - \frac{1}{e})^2 C_{OPT}$ , where  $C_{OPT}$  is the optimal value of the LP relaxation.*

**PROOF.** For every relay channel  $i$ , sort the users in the non-increasing order of  $\hat{w}_{ij}^{\hat{s}_j}$ , where  $s_j$  is the relay subset assigned to user  $j$  and  $i \in s_j$ . Assume that the sorted users are  $1, 2, \dots, K$  with  $\hat{w}_{i1}^{\hat{s}_j} \geq \hat{w}_{i2}^{\hat{s}_j} \geq \dots \geq \hat{w}_{iK}^{\hat{s}_j} \geq 0$  without loss of generality. The probability that channel  $i$  is assigned to user  $u$  in the rounded solution is atleast  $(\prod_{j=1}^{u-1} (1 - X_j^{\hat{s}_j})) X_u^{\hat{s}_j}$ . Thus, the expected contribution of relay channel  $i$  is given by  $\sum_{u=1}^K (\prod_{j=1}^{u-1} (1 - X_j^{\hat{s}_j})) X_u^{\hat{s}_j} \hat{w}_{iu}^{\hat{s}_j}$ . Using a lemma from [16] based on the arithmetic-geometric mean inequality, it can be shown that,

$$\sum_{u=1}^K (\prod_{j=1}^{u-1} (1 - X_j^{\hat{s}_j})) X_u^{\hat{s}_j} \hat{w}_{iu}^{\hat{s}_j} \geq \left(1 - (1 - \frac{1}{K})^K\right) \sum_{j, s_r} X_j^{\hat{s}_r} \hat{w}_{ij}^{\hat{s}_r}$$

Since the contribution of relay channel  $i$  in the optimal fractional solution is  $C_{i,:}^* = \sum_{j, s_r \in S_j} X_j^{\hat{s}_r} \hat{w}_{ij}^{\hat{s}_r}$ , we have the expected contribution of channel  $i$  in the rounded solution as,

$$E[\hat{C}_{i,:}^r] \geq \left(1 - (1 - \frac{1}{K})^K\right) C_{i,:}^* \quad (6)$$

Summing over all relay channels and applying the union bound, we now have the expected contribution of a user  $j$  to the objective function in the rounded solution of relay hop as,

$$E[\hat{C}_{:,j}^r] \geq \left(1 - (1 - \frac{1}{K})^K\right) \frac{C_{OPT}}{K} \quad (7)$$

where  $C_{OPT} = \sum_i C_{i,:}^*$ . The expected contribution (flow) of an access channel  $k$  conditioned on the flow from the relay hop can be obtained as,

$$E[\hat{C}_{k,:}^a | \hat{X}] \geq \left(1 - (1 - \frac{1}{K})^K\right) \sum_{j, s_a} Y_j^{\hat{s}_a} \hat{w}_{kj}^{\hat{s}_a} \cdot \frac{\hat{w}_j^{\bar{s}_r}}{\hat{w}_j^{\hat{s}_r}}$$

wherein the access weights are pessimistically updated to  $\bar{w}_{kj}^{\hat{s}_a}$  and  $\hat{X} = [\hat{X}_1^{\hat{s}_r}, \hat{X}_2^{\hat{s}_r}, \dots, \hat{X}_K^{\hat{s}_r}]$  is the vector of relay subset assignments. Unconditioning on the relay channel assignments, summing

over all access channels and interchanging summations, we have

$$E[\hat{C}^a] \geq \left(1 - (1 - \frac{1}{K})^K\right) \sum_{j=1}^K E[\hat{w}_j^{\bar{s}_r}]$$

where we have used  $E[\hat{C}_{k,:}^a] = E[E[\hat{C}_{k,:}^a | \hat{X}]]$  and  $\sum_{k, s_a} Y_j^{\hat{s}_a} \hat{w}_{kj}^{\hat{s}_a} = \hat{w}_j^{\hat{s}_r}$ . Further,  $E[\hat{w}_j^{\bar{s}_r}] = E[\hat{C}_{:,j}^r]$  and using equation 7, we have the expected net flow on second hop as,

$$E[\hat{C}] = E[\hat{C}^a] \geq \left(1 - (1 - \frac{1}{K})^K\right)^2 C_{OPT}$$

Since  $(1 - (1 - \frac{1}{K})^K) \geq (1 - \frac{1}{e} + \frac{1}{32K^2})$  we now have  $E[\hat{C}] \geq (1 - \frac{1}{e})^2 C_{OPT}$ .  $\square$

Now the question is whether  $C_{OPT}$  can be obtained. Since the LP relaxation of (3) has exponentially many variables we cannot apply a standard LP algorithm. However, one can write the corresponding dual formulation, which has a polynomial number of variables but an exponential number of constraints. The constraints can be equivalently represented as a polytope for every user,  $\mathcal{P}_j$ . Then we have the following lemma from [16].

**LEMMA 1.** *Given a polynomial time  $\beta$ -approximation separation algorithm for  $\mathcal{P}_j$ , one can obtain a  $(\beta - \delta)$  approximation for the dual problem and consequently for the primal problem, where  $\delta$  can be chosen to be exponentially small.*

In our case, the separation algorithm can be shown to be composed of two independent single user problems (single bin/knapsack problems), for which an FPTAS is known. Thus, we can obtain a polynomial-time  $(1 - \epsilon)$ -solution for the dual problem and consequently for the primal LP relaxation as well, where  $\epsilon > 0$  is an arbitrarily small constant. Putting all the results together, we have

**THEOREM 4.** *Algorithm DIV3 has an expected value of atleast  $(1 - \frac{1}{e})^2 (1 - \epsilon) C_{OPT}$  and hence provides an approximation guarantee of atleast  $(1 - \frac{1}{e})^2 - \epsilon$ , where  $\epsilon > 0$  is an arbitrarily small constant.*

### 3.3 Practical Significance of Algorithms

Going from the proposed matching algorithm for PCA (DIV1) to the two LP rounding based algorithms for FCA (DIV2, DIV3), the performance improves but so does the complexity. Since the algorithms have to be executed in real-time at the granularity of every scheduling epoch, trading off a slight degradation in performance for a significant reduction in complexity takes precedence. DIV3's iterative nature might not be conducive for real-time implementation in wireless systems. On the other hand, DIV2 and (more so) DIV1 are conducive for real-time implementation but then it is important to characterize their performance degradation. DIV2's performance degradation from rounding will be small when there are more sub-channels than the number of users in the system as captured in its approximation guarantee. The performance degradation in DIV1 arises from pair-wise restriction on channel assignments and from lack of incorporation of finite user buffers. Both these drawbacks are pronounced only when the number of channels is large compared to the number users. Thus, we find that while DIV3 provides a theoretically efficient solution, a load-adaptive combination of DIV1 and DIV2 provides an efficient practical solution, with DIV1 performing well for large number of users ( $\#users \geq \#channels$ ) and DIV2 for large number of channels ( $\#channels > \#users$ ). We substantiate these claims in the evaluations in Section 5.

## 4. SPATIAL REUSE SCHEDULING

Building on the diversity scheduler, we now design a scheduler for leveraging both diversity and spatial reuse to improve network capacity. Unlike diversity scheduling, where channels were not reused on the relay and access hops in tandem and hence a *flow schedule* (over two slots) was sufficient, the problem is now to find a *link schedule* that allows channels to be reused spatially on both the hops to maximize the aggregate marginal utility. In addition to the half duplex constraint of RS, MS operating on the access hop now have to incorporate interference from BS and RS operating on the relay hop have to incorporate interference from other RS operating on the access hop in tandem. For every slot, we have

$$S_{\max} = \arg \max_S \left\{ \sum_{j \in S} \beta_k \sum_{n=1}^N \sum_{h=1}^2 (\Delta U)_{n,j,h} I_{n,j,h} x_{r_{\text{relay}(j),h}} \right\}$$

$$\sum_{j=1}^K I_{n,j,h} x_{r_{\text{relay}(j),h}} \leq 1, \quad \forall n, h; \quad I_{n,j,h} = \{0, 1\}$$

$$x_{q,1} + x_{q,2} \leq 1, \quad \forall q; \quad x_{q,h} = \{0, 1\}$$

where  $I_{n,j,h}$  and  $x_{q,h}$  are binary functions indicating schedule of user  $j$  on channel  $n$  at hop  $h$ , and activation of relay  $q$  on hop  $h$  respectively. The first constraint indicates that channels are reused only across hops, while the second captures the half-duplex constraint of the RS. There arise several challenges in solving the above problem: (i) The marginal utility  $(\Delta U_{n,j,1})$  of a user  $j$  on channel  $n$  on hop 1, depends on its instantaneous relay channel rate, which in turn depends on the interference generated from the RS assigned to the same channel on hop 2, and hence on  $I_{n,j,2}$ . This results in non-linearity of the objective function, making the problem NP-hard [1]. (ii) Obtaining a *link* schedule requires the estimation of the independent *link* marginal utilities on the individual hops of the flow. However, the nature of the marginal utility of the flow does not allow decoupling into independent link (hop) components.

### 4.1 Spatial Reuse Algorithm: SR-DIV\*

We take a different approach in addressing the above challenges. Any schedule that enables spatial reuse will have a set of RS that will be scheduled on the relay hop and another (disjoint) set of RS that will be scheduled on the access hop in tandem on the same set of channels. Using this observation, our algorithm starts with an explicit partitioning of the RS. The essence of the algorithm can be described as follows: (i) BS (logically) partitions the set of RS into two disjoint sets,  $R_{RS}$  and  $A_{RS}$  representing the set of RS that will operate on the relay and access hops respectively in a given slot. (ii) BS runs one of our proposed diversity scheduler, DIV\*(1/2/3) on each of these sets to obtain two flow schedules. This not only retains the performance guarantees with respect to diversity exploitation, but also does not require the decoupling of the flow marginal utilities into their link components. (iii) The two flow schedules are not obtained independently, but are determined subject to the interference generated by each other. (iv) From the flow schedules obtained on the two disjoint sets of RS, a link schedule exploiting spatial reuse across the sets and diversity within the sets is generated. The algorithm is presented in Figure 3 and explained below.

#### 4.1.1 Partitioning

The goal is to find the optimal partition of the set of RS, such that the sum of the aggregate utilities of the flow schedules obtained on the two partitions is maximum. The problem can be shown to be NP-hard by giving a polynomial-time reduction from the *multiple knapsack problem*. The problem is made especially hard due to

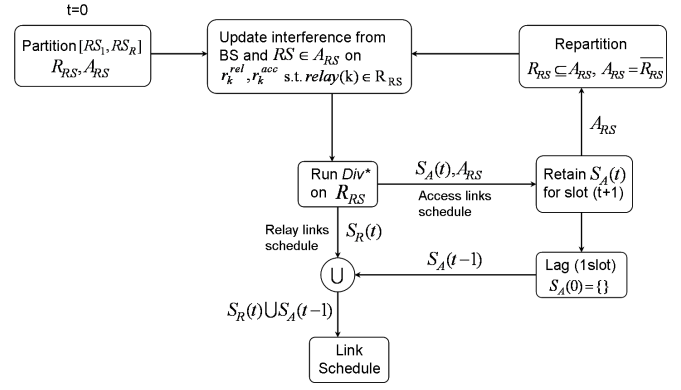


Figure 3: Algorithm

the *dependency* between the schedules obtained in the two partitions arising from interference. Given that the solution has to be run at the BS at the granularity of frames in real-time and since no polynomial-time solution is likely, we relax the problem to partitioning the set of RS based only on the traffic load in the network. This allows solving the relaxed problem in polynomial time, but to ensure that the sub-optimality in performance due to relaxation is kept minimal, the following constraints are incorporated. (i) The partitioned sets must be contiguous in order to keep the interference between the two sets minimal. This allows for negligible interference at RS away from the edge of the sets, with the interference at the edge of the sets accommodated through appropriate (orthogonal) channel assignments in the edges across sets. This allows for more flexibility in scheduling in the two sets, contributing to larger diversity and hence throughput. (ii) The partition size must ensure that the traffic load (users/flows) and diversity gains are balanced between the partitions to prevent under-utilization. Further, it must be automatically adapted by the scheduler at every frame based on perceived traffic load in the network. The relaxed partition problem with the above constraints is solved efficiently using the following dynamic program.

Given a set of RS, the objective is to partition the set into two contiguous sets such that the (QoS weighted) load between the two sets is balanced to the best extent possible. This is equivalent to minimizing the maximum (weighted) load over the two partitions. Since the two sets are completely defined by the starting and ending indices of one of the contiguous sets (partitions), let the maximum load of the partitioned sets be given by  $L[q, d]$ , where  $q, d \in [1, R]$  are the starting element index (RS) and the length of one of the partitions and  $R$  is the number of RS in the network. Note that, since the RS are placed in a circular geometry,  $q$  and  $d$  wrap around after  $R$ . Let  $w_q$  denote the load associated with  $RS_q$ ;  $\ell_{q,d}$  be the load associated with partition  $(q, d)$  consisting of  $[RS_q, RS_{q+d-1}]$ ; and  $W$  be the total load in the network. We have,

$$w_q = \sum_j \beta_j \cdot \mathbf{1}(j \in RS_q), \quad W = \sum_{q=1}^R w_q$$

The following dynamic program yields the desired partition.

$$(q, d)^* = \arg \min L[q, d]$$

$$L[q, d] = \max \{ \ell_{q,d}, W - \ell_{q,d} \}, \quad \forall (q, d)$$

$$\ell_{q,d} = \ell_{q-1,d+1} - \ell_{q-1,1}$$

The base cases from which the cost of the larger partitions can be

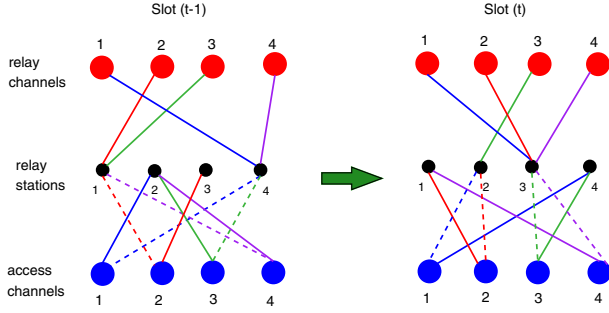


Figure 4: Illustration for exploiting spatial reuse and diversity

built are,

$$\ell_{1,d} = \sum_{q=1}^d w_q \quad \forall d, \quad \text{and} \quad \ell_{q,R} = W \quad \forall q$$

Thus, there are  $O(R^2)$  partitions. However, the cost of a larger partition is obtained in constant time using the cost of previously computed smaller partitions. Further, the partition yielding the minimum cost can also be kept track of in constant time at each step. Hence, the partitioning algorithm runs in  $O(R^2)$  time to yield the partitions with the best balanced load, as opposed to  $O(R^2 K)$  in a conventional approach.

#### 4.1.2 Schedule within Partitions

We use the topology in Figure 4 as a running example for illustration. Channels  $\{1, 2, 3, 4\}$  on the relay and access hops are available to be scheduled to four users, associated with relays  $\{1, 2, 3, 4\}$ , with one user per relay. Let the RS operating on the relay and access hops in slot  $(t-1)$  be  $R_{RS} = \{RS_1, RS_4\}$  and  $A_{RS} = \{RS_2, RS_3\}$ . The access links of the flows  $(S_A(t-1))$ , dashed lines in slot  $t-1$ , whose relay links were scheduled from  $R_{RS}$  in slot  $(t-1)$  constitute the access hop schedule for the next slot  $t$  (solid lines on access hop in slot  $t$ ). The new  $R_{RS}$  for the next slot  $t$  is chosen to be a **subset** of the existing access set, ( $R_{RS} \subseteq A_{RS} = \{RS_2, RS_3\}$ ) by applying our partition algorithm to  $A_{RS}$  to allow for load balancing (repartitioning) between the partitions in response to varying traffic conditions.  $A_{RS}$  is then updated to the complement of  $R_{RS}$ , namely  $\overline{R_{RS}} (\{RS_1, RS_4\})$ . Our diversity scheduler (DIV\*) is then run on the new  $R_{RS}$ , taking into account the interference generated from  $A_{RS}$ . From the resulting diversity flow schedule, the relay links ( $S_R(t)$ , solid lines in slot  $t$  at  $RS_2, RS_3$ ) are scheduled in tandem with the access links waiting from the previous flow schedule ( $S_A(t-1)$ , dashed lines from slot  $t-1$  at  $RS_1, RS_4$ ), thereby generating a link schedule that exploits spatial reuse. The access links from the current flow schedule ( $S_A(t)$ , dashed lines in slot  $t$  at  $RS_2, RS_3$ ) are retained for schedule in the next slot and the process repeats.

#### 4.1.3 Incorporation of Interference

Before applying DIV\*, the instantaneous rates fed back from MS and RS must incorporate interference. For any MS, the source of interference (BS) does not change and there is no power adaptation across channels. Thus, the MS can directly incorporate the interference from BS ( $\chi_{BS \rightarrow k,n}$ ) in their instantaneous access rate feedback without (a priori) knowledge of the specific relay link to be scheduled on the same channel:  $r_{k,n}^{acc} = \log(1 + \frac{P_{k,n}}{N_{k,n} + \chi_{BS \rightarrow k,n}})$ ,  $\forall n$ , where  $P_{k,n}$  and  $N_{k,n}$  correspond to the received signal and noise power at MS  $k$  from its associated RS. The RS  $\in R_{RS}$

operating on the relay hop will experience interference from RS  $\in A_{RS}$ . However, the BS is already aware of the access hop schedule ( $S_A(t-1)$ ) in slot  $t$ , one slot prior to their actual schedule. Hence, this information is conveyed by BS to the RS  $\in R_{RS}$  in the form of a bitmap ( $BM_{acc}$ ) broadcast. The anticipated interference ( $\chi_{j \rightarrow q,n}$ ) from  $RS_j \in A_{RS}$  at  $RS_q \in R_{RS}$  ( $\chi_{j \rightarrow q,n}$ ) is then incorporated in the relay channel feedback as,  $r_{q,n}^{rel} = \log(1 + \frac{P_{q,n}}{N_{q,n} + \sum_{j \in Access\_RS} \chi_{j \rightarrow q,n} B_{j,n}})$ ,  $\forall n$ , where  $B_{j,n} = 1$  if  $BM_{acc}(n) = RS_j$ , and 0 otherwise.  $P_{q,n}$  and  $N_{q,n}$  correspond to the received signal and noise power at RS  $q$  from BS.

An interference-aware DIV\* coupled with the partitioning mechanism forms the core of the algorithm that helps construct link schedules in *polynomial time* from two interference-dependent flow schedules without requiring the decoupling of hop marginal utilities. The sub-optimality of the solution arises from the (i) relaxation of interference dependency in the partitioning process, as well as in (ii) obtaining the relay hop schedule subject to an access hop schedule obtained instead of jointly optimizing them. However, the careful partitioning of the RS along with our efficient diversity scheduler, allows for sufficient diversity gains in the two sets, which in turn helps keep the sub-optimality of the spatial reuse schedule low, notwithstanding its low running time complexity. This is evident in the evaluations where the scheduler performs reasonably close to the upper bound.

**Spatial reuse within access links:** Spatial reuse within access links was not exploited since it comes at the cost of reuse across hops. Further, even if leveraged, it would not translate to improved performance unless the access hop forms the network bottleneck. However, it can be easily integrated into our diversity solutions (DIV2, DIV3), wherein the constraint of assigning at most one user to an access channel, will be transformed to assigning at most one user *per contention region* to an access channel. A contention region can be defined with respect to every RS, which defines the set of RS (on either side of the considered RS) that cause interference to its associated MS when operating on the same channel. This increases the number of constraints on the access channels, and hence the running time by a factor of  $R$  (# relays) but the solutions and performance guarantees would still apply.

## 5. PERFORMANCE EVALUATION

An event-driven packet level simulator written in C++, named QNS [17] coupled with the GNU LP kit is considered for evaluation of the proposed solutions. A single cell relay-enabled OFDMA downlink system is considered. The extended radius of the cell is assumed to be about 600m. RS are distributed uniformly within a region of  $250m \leq r \leq 350m$ . The wireless links incorporate path loss and Rayleigh fading as well as interference from other links operating on the same channel. Each user's Rayleigh channel has a Doppler fading equivalent to a velocity of 3-10 Km/hour. We consider constant bit rate (CBR) applications as the generators of traffic. A time slot is considered to be of 5 ms duration, and carrier frequency is assumed to be 2 GHz. The peak rate of the individual sub-channels is 250 Kbps. The number of users, relays, sub-channels and buffer sizes are the parameters of variation. The data flows are sent at 125 Kbps. We consider traffic loads ranging from low to high by varying the number of users (flows) in the system. The metrics of evaluation are throughput and utility.

### 5.1 Performance of Diversity Algorithms

We first evaluate the two algorithms for exploiting diversity, namely DIV1 and DIV2 and compare them against the optimal LP fractional solution. The user weights used in the LP's take into account

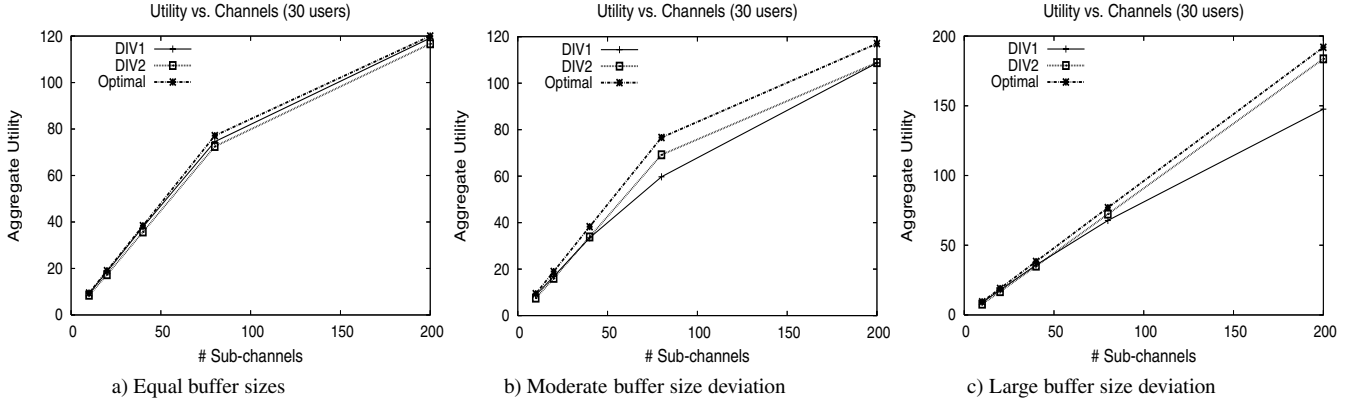


Figure 5: Diversity exploitation as function of user buffer size deviation.

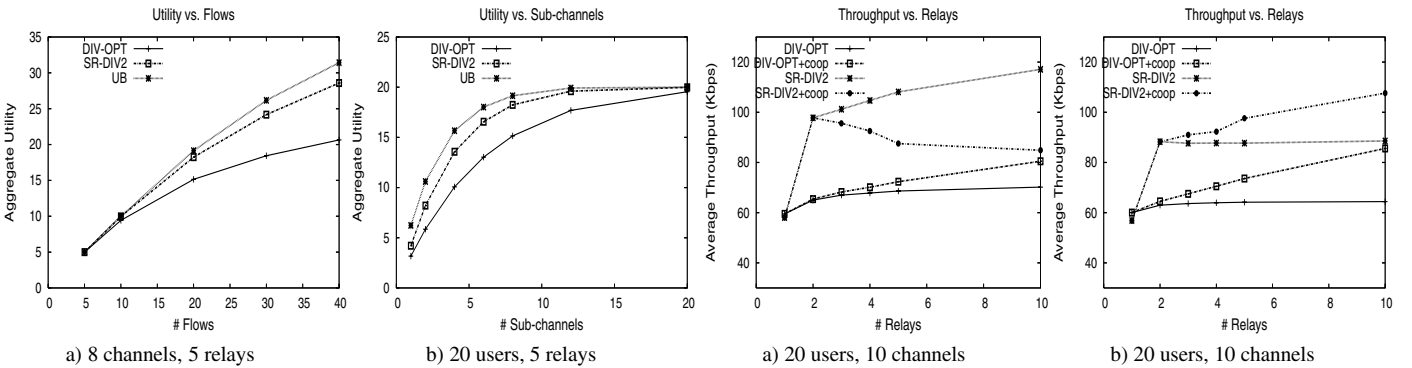


Figure 6: Exploiting diversity and spatial reuse.

Figure 7: Impact of Cooperative Diversity.

both the user's throughput and Rayleigh fading channel gain on the specific channel considered. The aggregate utility results are presented in Figure 5 for 30 users and 5 relays in the system. In Figure 5(a) both DIV1 and DIV2 perform very close to the optimal. This trend is preserved as long as the buffer sizes are the same across users, irrespective of the absolute size. However, when the deviation in the buffer size across users is increased in Figures 5(b) and (c), the degradation starts to increase. DIV1's degradation is small with smaller number of channels but increases to about 25% when the number of channels is large as expected. DIV2 on the other hand, maintains a constant performance gap from the optimal, resulting in the loss percentage decreasing with increasing channels to about only 10%. Given the reduced complexity of DIV1 over DIV2, DIV1 can be used as an efficient algorithm for exploiting diversity at small-moderate number of channels, while DIV2 can be employed for large number of channels, thereby corroborating our claims.

## 5.2 Performance of Spatial Reuse Algorithm

We now evaluate the performance of SR-DIV\* (with DIV2 as the diversity scheduler) against that of optimal diversity (DIV-OPT) and a loose upper bound (UB). For upper bound, we assume that the capacity on the relay links is achievable through a genie. This is incorporated by allowing the two-hop flow rates of the scheduled users to correspond directly to their one-hop relay link rates in the absence of interference. The aggregate utility results for SR-DIV2 are presented in Figure 6(a) and (b) as a function of increasing users and channels in the system respectively. For a given

network capacity, when the number of users (channels) is small (large), the injected traffic load can be sustained using diversity gains alone. However, when the number of users (channels) increases (decreases), spatial reuse must be exploited to sustain a larger fraction of the injected traffic load. This in turn results in the throughput gain of spatial reuse increasing to over 50% over diversity. The gain then stabilizes with increasing load once the available spatial reuse has been exploited. The results clearly indicate SR-DIV2's superiority in both utilization and fairness resulting from spatial reuse.

In addition to the significant gains over DIV-OPT, SR-DIV2 also performs reasonably close to the (loose) upper bound with a maximum deviation of about 20%. This is especially noteworthy, given that the optimal is going to be lower than the upper bound. Further, it also indicates that the additional gain that can result from a further degree of spatial reuse exploitation within access links is not appreciable.

## 5.3 Impact of Cooperative Diversity

Figure 7 captures the impact of cooperative diversity. In Figure 7(a) the capacity of the relay and access channels are the same and neither hop is a clear bottleneck. However, in Figure 7(b) the relay channel capacity is made 50% more than the access channel, making the access hop a clear bottleneck, where increasing relays increases spatial reuse gains. Addition of cooperative diversity mechanisms improves performance for the diversity solutions as expected in all cases. However, they improve performance for the spatial reuse solution only in the access bottleneck case (b) and

in fact degrade performance in the general case (a) when number of relays increases. This is because, cooperative diversity mechanisms improve SNR at the MS receiver by adding more cooperating RS transmitters. However, this also increases interference on the RS operating on the relay hop in spatial reuse solutions. *The cooperative diversity gain, which improves access hop performance, outweighs the impact of interference and hence spatial reuse when the access hop forms the network bottleneck. However, it degrades performance in the general network case.*

## 6. DISCUSSIONS

**Multicast flows:** Our diversity solutions also apply to multicast flows with the same performance guarantees as for unicast flows. However, the scheduling model must be appropriately adapted to multicast groups (sessions) with the instantaneous session rate determined by the bottleneck user in the group. While wireless broadcast is a key enabler for delivering multicast data efficiently, identical data sent on the same access channel but from different RS could still interfere at a MS due to channel fading, thereby necessitating different cooperative diversity models. Further, since spatial reuse comes at the cost of wireless broadcast advantage, the relative significance of diversity and spatial reuse gains will be different for multicast compared to unicast.

**Power control:** It has been shown that most of the diversity gains can be leveraged through channel-dependent scheduling alone, while power control across channels provides only marginal additional gains in OFDMA cellular systems [13]. Similarly, in the two-hop model, while power control can help reduce the interference generated during spatial reuse across hops, if such interference can be taken care of through carefully generated schedules, then the gains from power control will once again be marginal.

**Synchronization and Overhead:** In current 802.16j relay systems, the transmissions on the two hops are synchronized on a frame-basis, allowing efficient allocation and interference management of sub-channels on the two hops. Further, a *sounding* mechanism is provided for BS to collect channel and interference information from RS and MS. This results in an overhead of  $O((K + R^2)N)$ . However, one can leverage a priori knowledge of interference and some mechanisms for feedback consolidation (from MS) at RS to reduce the feedback sent to BS to only  $O(RN)$  without any appreciable degradation in performance. This makes the overhead tractable for a two-hop network in the absence of power control. However, in the presence of power control, one cannot escape the overhead of  $O((K + R^2)N)$ , which may not justify the marginal gains resulting from power control and hence requires further investigation.

**Multiple hops:** While synchronization and overhead are tractable for a two-hop network, they may not be feasible for multiple hops, in which case one would have to resort to more decentralized approaches. However, most of the envisioned applications for relay systems fall under the two-hop category, thereby emphasizing the benefits of the proposed centralized solutions.

## 7. CONCLUSIONS

We have considered the specific problem of scheduling user traffic on the multiple OFDM sub-channels over the two hops of the relay-enabled wireless networks. We proposed scheduling algorithms that help leverage the diversity and spatial reuse gains from these networks. We showed that even the scheduling problem to exploit diversity gains alone is NP-hard and provided both theoretically and practically efficient polynomial-time algorithms with approximation guarantees. We also proposed an efficient polynomial-

time scheduling algorithm for exploiting both spatial reuse as well as diversity. Evaluations of the proposed solutions highlighted the relative significance of various diversity and spatial reuse gains with respect to varying network conditions.

## 8. REFERENCES

- [1] G. Brar, D. Blough, and P. Santi, "Computationally efficient scheduling with the physical interference model for throughput improvement in wireless mesh networks," in *ACM MOBICOM*, 2006.
- [2] T.-S. Kim, H. Lim, and J. C. Hou, "Improving spatial reuse through tuning transmit power, carrier sense threshold, and data rate in multihop wireless networks," in *ACM MOBICOM*, Sept 2006.
- [3] Z. Zhang, Y. He, and K. P. Chong, "Opportunistic downlink scheduling for multiuser ofdm systems," in *IEEE WCNC*, Mar 2005.
- [4] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks - Part I: Theoretical Framework," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, Mar 2005.
- [5] M. Andrews and L. Zhang, "Scheduling algorithms for multi-carrier wireless data systems," in *ACM MOBICOM*, Sept 2007.
- [6] A. So and B. Liang, "Effect of relaying on capacity improvement in wireless local area networks," in *IEEE WCNC*, Mar 2005.
- [7] S. Mengesha and H. Karl, "Relay routing and scheduling for capacity improvement in cellular w lans," in *WiOpt*.
- [8] C. Hoymann, P. Dallas, A. Valkanas, A. Gosteau, D. Noguét, and R. Hoshyar, "Flexible relay wireless ofdm-based networks," in *Funded by European Commission*, 2006.
- [9] N. Challa and H. Cam, "Cost-aware downlink scheduling of shared channels for cellular networks with relays," in *IEEE International Conference on Performance, Computing, and Communications*, 2004.
- [10] H. Viswanathan and S. Mukherjee, "Performance of cellular networks with relays and centralized scheduling," *IEEE Transactions on Wireless Communications*, vol. 4, no. 5, Sep 2005.
- [11] M. Herdin, "A chunk based ofdm amplify-and-forward relaying scheme for 4g mobile radio systems," in *IEEE ICC*, Jun 2006.
- [12] A. Hottinen and T. Heikkinen, "Subchannel assignment in ofdm relay nodes," in *Proc. of CISS*, Mar 2006.
- [13] J. Jang and K. B. Lee, "Transmit power adaptation for multi-user OFDM systems," *IEEE JSAC*, vol. 21, no. 2, pp. 171–179, 2003.
- [14] B. Radunovic and J. Le Boudec, "Rate performance objectives of multi-hop wireless networks," in *IEEE INFOCOM*, Mar 2004.
- [15] C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity (Chapter 11)*, Prentice Hall, 1982.
- [16] L. Fleischer, M. Goemans, V. Mirrokni, and M. Sviridenko, "Tight approximation algorithms for maximum general assignment problems," in *ACM SODA*, 2006, pp. 611–620.
- [17] R. G. Mukthar, "Qns: Queuing network simulator," in *QNS v0.1*, <http://www.cubinlab.ee.mu.oz.au/rgmukht/qns>, Nov 2003.