WiMAX Relay Networks: Opportunistic Scheduling to Exploit Multiuser Diversity and Frequency Selectivity

Supratim Deb
Bell Labs India, Bangalore
supratim@alcatel-lucent.com

Vivek Mhatre*
Motorola, Arlington Heights, IL
vivekmhatre@motorola.com

Venkatesh Ramaiyan†
Indian Institute of Science, Bangalore
rvenkat@ece.iisc.ernet.in

ABSTRACT

We study the problem of scheduling in OFDMA-based relay networks with emphasis on IEEE 802.16j based WiMAX relay networks. In such networks, in addition to a base station, multiple relay stations are used for enhancing the throughput, and/or improving the range of the base station. We solve the problem of MAC scheduling in such networks so as to serve the mobiles in a fair manner while exploiting the multiuser diversity, as well as the frequency selectivity of the wireless channel. The scheduling-resources consist of tiles in a two-dimensional scheduling frame with time slots along one axis, and frequency bands or sub-channels along the other axis. The resource allocation problem has to be solved once every scheduling frame which is about $5-10$ ms long. While the original scheduling problem is computationally complex, we provide an easy-to-compute upper bound on the optimum. We also propose three fast heuristic algorithms that perform close to the optimum (within 99.5%), and outperform other algorithms such as OFDM^{2}A proposed in the past. Through extensive simulation results, we demonstrate the benefits of relaying in throughput enhancement (an improvement in the median throughput of about 25%), and feasibility of range extension (for *e.g.*, 7 relays can be used to extend the cell-radius by 60% but mean throughput reduces by 36%). Our algorithms are easy to implement, and have an average running time of less than 0.05 ms making them appropriate for WiMAX relay networks.

Categories and Subject Descriptors: C.2.0 [General]: Data Communications; C.2.1 [Computer Communication Networks]: Network Architecture and Design-*Wireless communication*

General Terms: Algorithms, Performance

1. INTRODUCTION

IEEE 802.16e, popularly known as WiMAX, is the fourth generation (4G) standard for broadband wireless access [9]. WiMAX uses large chunks of spectrum (10-20 MHz or more), and delivers high bandwidth (up to 75 Mbps). The physical layer of WiMAX uses scalable-OFDMA since OFDM has twofold benefits in terms of robustness to multi-path fading, and ease of DSP implementation due to the use of the FFT algorithm.

Despite the high bandwidth promised by WiMAX, there are several issues that network operators face during actual deployment of these networks. The first problem is that of *dead spots* or *coverage holes*. Such spots of poor connectivity are formed due to high path-loss, and shadowing due to obstacles such as large buildings, trees, tunnels, etc. and this leads to degradation in overall system throughput. The other key design challenge is that of *range extension*. At times, it is required to provide wireless connectivity to an isolated area outside the reach of the nearest base station (BS). The above problems of throughput enhancement by filling coverage holes and range extension can be easily tackled by deploying additional base stations. However, such a solution could be an overkill, and too expensive in several scenarios. In such contexts, relay stations are a cost-effective alternative. Relay stations (RS) act as MAC-layer repeaters to extend the range of the base station. An RS decodes and forwards MAC-layer segments unlike a traditional repeater which merely amplifies and retransmits PHY-layer signals. Hence, an RS may use a different modulation coding scheme for reception and forwarding of a MAC segment. Although more advanced RS designs in which multiple RSs cooperate in a network-MIMO configuration are possible, we only focus on decode-and-forward RSs. The IEEE 802.16j task-group [10] has been formed to extend the scope of IEEE 802.16e to support mobile multi-hop relay (MMR) networks.

Unlike a BS, an RS has a significantly simpler hardware and software architecture, and hence lower cost. An RS merely acts as a link layer repeater, and therefore does not require a wired backhaul. Furthermore, an RS need not perform complex operations, such as connection management, hand-offs, scheduling, etc. Also, an RS typically operates at much lower transmit power, and simply requires lower-MAC and PHY layer stack. All these factors lead to much lower cost of an RS, and thus, relay networks are evolving as a low-cost option to fill coverage holes and extend range in many scenarios. Although, conceptually, the relay networks are simple, there are several design challenges involved: (i) *Relay placement* [24], (ii) *Scheduling mobiles and relays over time and frequency* [18], (iii) *Inter-relay hand-off* [10], (iv) *Routing to/from the BS* [4, 18, 17]. The focus of this paper is primarily on *scheduling*. Although we only address downlink scheduling, all our algorithms can be easily extended to the uplink scenario.

Since WiMAX networks typically operate over 10-20 MHz or wider bandwidth, there is a significant amount of frequency selectivity over all the links [20]. In other words, if the operating spectrum is subdivided into narrow sub-channels (as is done in OFDMA), there is substantial variation in the rates that can be supported over each sub-channel of a given link. Scheduling in WiMAX relay network is carried out by BS that computes and broadcasts the schedule once in a *scheduling frame* where a schedul-

*This work was carried out when Vivek Mhatre was working with Bell Labs India.

†Part of this work was carried out when Venkatesh Ramaiyan was visiting Bell Labs India.

MobiCom'08, September 14–19, 2008, San Francisco, California, USA.

ing frame consists of multiple time-slots. Thus, the set of available resources in a WiMAX like OFDMA networks can be conceived as tiles in a two-dimensional tiling structure consisting of time slots along one axis, and sub-channels along the other axis. Thus, the scheduling problem in OFDMA relay networks is the problem of assigning transmission opportunities (tiles) to each link in the network to maximize a certain objective function. This time $\times$ frequency tiling problem is at the heart of most OFDMA scheduling problems [3]. In relay networks, there are additional constraints due to synchronization in a multi-hop topology, use of a single transceiver at the relays, and flow conservation due to multi-hop relaying. Finally, the scheduling decisions in WiMAX networks have to be made in a timely fashion since the schedule is typically disseminated once every $5 - 10$ ms which is the typical order of magnitude of coherence time of the channel[20]. Thus, the problem of scheduling for fair-rate allocation in WiMAX relay networks poses several unique challenges. We highlight some of these challenges through a simple illustration:

Example. *Consider a simplistic relay network with one BS, one RS, and three mobiles. Mobiles M_1 and M_2 communicate to the BS via the RS in a two-hop fashion, and mobile M_3 communicates directly with the BS. Suppose the available spectrum is subdivided into two OFDM sub-channels, C_1 and C_2. Suppose, there are* **seven** *time-slots in a scheduling frame. We wish to decide, who transmits to whom at which time slot, and over which sub-channel during a given scheduling frame. Let $r_i(M_j, RS)$ be the rate between M_j and RS over sub-channel indexed i, and the other rates are denoted similarly. To bring out the effect of frequency selectivity of the wireless channel, we specify different rates (in bits/time-slot) as follows.*

$$r_1(M_1, RS) = r_2(M_2, RS) = 200$$
$$r_1(M_2, RS) = r_2(M_1, RS) = 50$$
$$r_1(RS, BS) = r_2(RS, BS) = 50$$
$$r_1(M_3, BS) = r_2(M_3, BS) = 26$$

Suppose there is a large backlog of data for each mobile at the BS in the current scheduling frame. Now, the problem can be stated as follows: Given the preceding system, find time-slot and sub-channel assignment for downlink transmission to the mobiles and the relays such that, (i) the total data transmitted is maximized[1], (ii) the BS and the RS, being in the same cell-sector, do not transmit simultaneously in a slot, and (iii) the RS does not transmit and receive concurrently (even over different sub-channels) as each RS has a single transceiver. *One simple solution is to transmit to M_3 over the seven slots over both the sub-channels, giving a total data transmission of $52 \times 7 = 364$ bits in the scheduling-frame. However, this can be improved by the following solution: the BS transmits to RS over the first 4 slots over both sub-channels (transmitting 400 bits in total), during slot 5 RS transmits to M_1 over sub-channel C_1 and transmits to M_2 over C_2, during slots 6 and 7 the BS transmits to M_3 over both the sub-channels. This solution gives a total data transmission of $400 + 52 \times 2 = 504$ bits, thereby improving upon our first solution by nearly 40%. Note that, if $r_1(M_3, BS) = r_2(M_3, BS) = 40$ instead, it would be optimal to transmit to M_3 over all the seven slots.* □

A moment's reflection shows that the above problem is much easier in the absence of any relays and the solution is very simple: over each sub-channel, transmit to the mobile with the best rate. However, with multiple relays, a large number of sub-channels, multiple hops, and due to discrete nature of the available number of sub-channels and time slots, the combinatorial complexity of the solution space blows up. Thus, in order to exploit the diversity of wireless channels across frequency and across different users/links, each scheduling decision involves solving a combinatorially hard problem. Furthermore, since the MAC resource allocation has to be done in real-time (typically within a $5 - 10$ ms scheduling frame), we need low-complexity and efficient MAC algorithms to perform the resource allocation.

Although there is rich literature on opportunistic scheduling in single hop OFDM networks [5, 14, 7], and single hop narrow-band cellular networks [23], these works cannot be extended to the multi-hop relay scenario where each hop has potentially different rates over different sub-channels. Scheduling in OFDMA relay networks has been recently addressed in [4], but this does not exploit the frequency-selectivity of the wireless channel. Closest to our work is the work in [18], where the authors propose a scheme termed as OFDM^{2}A that accounts for frequency-selectivity and provides significant gains over round-robin scheduling. In OFDM^{2}A, sub-channel allocation is done on a per mobile basis, *i.e.*, each mobile is allocated a sub-channel which is used by all the intermediate links of the relay network during the entire scheduling-frame. However, we observe that the above approach of allocating sub-channels on a per-mobile basis for the entire path has the following fundamental limitation. Consider two mobiles i_1 and i_2 whose path to the BS passes through a common relay u, and the mobiles are assigned sub-channels j_1 and j_2 respectively. Depending on the rates of links associated with u over sub-channels over j_1 and j_2, node u may be required to transmit data intended for mobile i_1 over sub-channel j_1, while it is concurrently receiving data intended for mobile i_2 over sub-channel j_2. Thus, implementing OFDM^{2}A clearly requires the relays to be equipped with multiple-radios, and does not comply with the IEEE 802.16j requirements. Our work does not have such limitations, and also outperforms OFDM^{2}A as we show in our results.

1.1 Main Contributions

The main contributions of this paper are as follows.

1. *Fair scheduling in OFDMA relay networks:* We develop a framework for proportional-fair scheduling in OFDMA-based relay networks in general, and IEEE 802.16j based WiMAX relay networks in particular. Our framework exploits multiuser diversity along with frequency-selectivity of the wireless channels. We show that the original tile scheduling problem is NP-hard, and is hard to approximate. We then provide an easy-to-compute performance upper bound using a relaxed LP. We use this upper bound as a benchmark for comparing the performance of our proposed algorithms.
2. *Low complexity MAC scheduling algorithms:* We propose several low-complexity novel scheduling algorithms that can be implemented in real-time in WiMAX relay networks. Extensive simulations demonstrate that our algorithms perform close to optimum (within 99.5% of the optimum). Our proposed scheduling algorithms have an average running time of less than 0.05 ms, and are therefore suitable for typical WiMAX scheduling frame durations of $5 - 10$ ms. Our proposed algorithms outperform OFDM^{2}A [18] which requires the relay nodes to be equipped with multiple transceivers in order to satisfy the synchronization constraints.
3. *Throughput enhancement using relays:* We evaluate the throughput benefit of WiMAX relay networks through detailed simulations. Our simulations suggest that, even a handful (three) of relays can improve the median throughput of the mobiles by up to 25%, and the mean throughput by up to 15% in a typical 1km sector.
4. *Range extension using relays:* We evaluate the performance of relays for range extension through comprehensive simulation experiments. We show that, compared to a no-relay scenario (only BS) in a cell with 1 km radius, 5 relays can be used to extend the cell-radius by 20% but with a mean throughput reduction of

[1] In this example, we attempt to maximize throughput simply for ease of illustration. A more desirable goal is to maximize a metric that also ensures some kind of *fairness* and we do account for that in the later sections.

11%, and 7 relays can be used to extend the cell-radius by 60% but with a mean throughput reduction of 36%.

The paper is structured as follows. Section 2 provides some background on WiMAX and proportional-fair schedulers and Section 3 describes our network model. Section 4 formulates the scheduling problem in generic OFDMA-based relay networks and develops a low-complexity scheduler. In Section 5, we develop schedulers for IEEE 802.16j based relay networks which are a special case of the generic OFDMA-based relay networks. Section 6 provides simulation results to demonstrate the performance of our algorithms, and the benefits of using relays. Related work is discussed in Section 7. Finally, we conclude in Section 8.

2. BACKGROUND

2.1 WiMAX

In the following, we discuss some key features of WiMAX. For a more detailed description, we refer the reader to [9, 10].

Sub-carrier permutation and sub-channels: Sub-channels in WiMAX consist of narrow frequency bands called sub-carriers. Sub-channelization can be done in two modes [9]. In the *diversity permutation mode*, sub-carriers that form each sub-channel are chosen randomly from the entire frequency spectrum. In the *contiguous permutation mode*, a sub-channel is made up of adjacent sub-carriers. Contiguous permutation is useful when mobiles are fixed or moving at a low-speed, since rate adaptation can be used to exploit frequency selectivity. The diversity permutation mode has been recommended for highly mobile users, or for users with very low signal-to-interference-plus-noise ratio (SINR) to exploit frequency diversity. In this work, we will assume contiguous permutation to exploit frequency selectivity of the wireless channel.

Modulation coding schemes and cross layer adaption in WiMAX: WiMAX allows three choices of modulation, namely, QPSK, 16-QAM, and 64-QAM along with three choices of FEC schemes for a total of six distinct permissible modulation-coding combinations. Depending on the measured SINR of a link, one of these six modulation-coding schemes can be used.

2.2 Proportional Fair (PF) Scheduling

In this section, we provide a brief primer on proportional fair scheduling, since our scheduling algorithms aim to provide proportional fairness. We start by giving the definition [13, 2] of such a notion of fairness.

Definition: A set of rates R_i is said to be *proportional fair (PF)* if R_i's are feasible, and if, for every other feasible set of rates S_i, the following holds:

$$\sum_i \frac{S_i - R_i}{R_i} \leq 0.$$

It can be shown that, if R_i's are proportional-fair, then R_i's also maximize the so called *proportional-fair metric* $\sum_i \log R_i$ over all possible feasible long-run rates. This also provides an equivalent definition of proportional-fair rate [13].

Achieving proportional-fair rates: PF rate allocation has been adopted as the notion of fair-rate allocation in many systems, particularly wireless systems like EVDO [23].

In the following, we provide conditions under which short-term rates converges to PF over long-run [15]. Let $d_i(t)$ be the data rate to mobile i at time-slot t, and let $R_i(t)$ be the average rate of mobile over the time horizon $[1, t]$, *i.e.*, $R_i(t) = \sum_{s=1}^{t} d_i(s)/t$. If $d_i(t)$'s maximize $\sum_i d_i(t)/R_i(t-1)$ among all feasible rate-vectors $d_i(t)$ for all t, then the long-run rates $R_i(t)$'s are proportionally-fair. Even if we replace the $R_i(t)$'s by exponentially smoothed average [2], the resulting $R_i(t)$'s converge to PF allocation as t becomes large [15]. *We note that the $R_i(t)$ in the denominator ensures that any mobile cannot be starved for a long duration.*

[2]In such an averaging R_i's are updated by $R_i(t) \leftarrow \alpha R_i(t-1) + (1-\alpha) d_i(t)$, where $\alpha < 1$ is a constant typically close to one.

Thus, we have a recipe for achieving PF allocation of rates [2]: *For all times t, assign data rates $d_i(t)$ such that $\sum_i d_i(t)/R_i(t-1)$ is maximized among all feasible d_i's.* However, this maximization problem can be challenging for two reasons. Firstly, the maximization problem could be combinatorial in nature, thus leading to a computationally hard problem. Secondly, the maximization has to be performed in real-time, for example, in WiMAX based networks, each scheduling decision has be performed in 5 ms. Thus, the solutions should be computationally very light. One of the goals of this paper is to solve the above mentioned maximization problem under the constraints imposed by OFDMA based relay networks.

3. NETWORK MODEL AND NOTATIONS

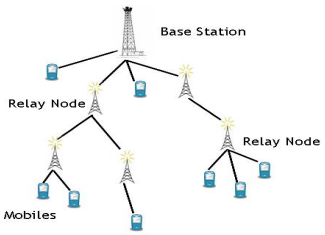


Figure 1: A relay network.

A WiMAX network is divided into cells (similar to cellular systems) and each cell is further divided into three 120° sectors. A WiMAX base-station performs MAC resource allocation separately for each sector in a cell. Our network model is that of a single sector in a WiMAX cell, consisting of a single base station node (denoted by BS, and node index 0) and multiple relay stations as shown in Figure 1. The BS is at the root of the tree, the RS's are the intermediate nodes of the tree and the mobile nodes are the leaf nodes of the tree as shown in the figure. The routing tree can be determined using link metrics such as Expected Transmission Time (ETT) [19]. When co-operative communication [20] is used by the RS's and the BS, our scheduling framework can be easily modified to take into account the combining gains.

Table 1 summarizes the notations used in the paper, $\mathcal{M}$ denotes the set of mobiles, $\mathcal{R}$ denotes the set of relays, and $\mathcal{R}^+$ denotes the set of relays and the BS. We use the convention that a set is denoted using script font, and its size is denoted using the normal font. Thus, the size of set $\mathcal{M}$ is M. Since the network has a tree topology, there are $M + R$ links in the network. Let $\mathcal{L}$ be the set of these links. For each node u, denote by l_u to be the link between u and its parent node in the tree, and denote by L_u the set of links between u and its children. Denote by $p(u)$ the parent of node u in the routing tree.

A scheduling frame consists of time-slots and sub-channels[3]. A time-slot and sub-channel combination (referred to as a *tile* in the rest of the paper) is the minimum allocable resource unit in the WiMAX MAC. Let T and τ respectively be the frame duration and time-slot duration in seconds. Each frame has $N = T/\tau$ time slots. The set of sub-channels are denoted by $\mathcal{C}$.

In every scheduling-frame, the BS computes and broadcasts the schedule for the entire cell-sector. Let $r_{(l,c)}(t)$ be the rate, in bits/time-slot, that can be supported on link l over sub-channel c in frame t. In this paper, we mostly work on developing algorithms for a given frame. Hence we drop the dependence on the frame index t from $r_{(l,c)}(t)$ and simply use $r_{(l,c)}$ to denote the channel rates in the frame under consideration.

Properties of system model: Our system model also has the following properties arising from practical considerations.

(P1) The rates $r_{(l,c)}$ at link l for sub-channel c are known to the BS at the beginning of every frame. The IEEE 802.16j standard has

[3]The terms *channel* and *sub-channel* are used interchangeably. Both refer to a sub-channel.

Notation	Description
$\mathcal{M}$	Set of Mobiles
$\mathcal{R}$	Set of Relays
$\mathcal{R}^+$	Set of Relays and BS
$\mathcal{C}$	Set of sub-channels
$\mathcal{L}$	Set of all the links
$\mathcal{T}_u$	Subtree rooted at node u
P_u	Set of links on the path from node u to the BS
S_u	Set of mobiles directly attached to relay/BS u
L_u	Set of links between $u \in \mathcal{R}^+$ and mobiles in S_u
$L_u^{'}$	Set of links between $u \in \mathcal{R}^+$ and its child relay nodes
l_i	Link between node i and its parent
$p(i)$	Parent of node i
L_u	Child links of relay node u
N	Number of slots in a frame
$r_{(l,c)}$	Rate of link l over sub-channel c in bits/slot
R_i	Long term average throughput of mobile i in bits/scheduling-frame

Table 1: Notations used.

specified methods for doing this [10].

(P2) There is no spatial reuse within a sector, *i.e.*, a sub-channel c is used only at one link at a time in a given sector. This is essential as the relays will lie within a sector of the same cell. Also, as we will see later in the paper, this is automatically ensured by the specifications of IEEE 802.16j standards.

(P3) From an architectural point of view [10], the mobiles are agnostic to the presence of relays, and there is no network layer communication between the relays and the mobiles. As a result, relays act simply as MAC-layer repeaters (unlike mesh routers which can queue and forward packets). We therefore assume that packets are not queued at the relay nodes, *i.e.*, the flow constraints are strictly met over each frame duration at each relay node[4].

(P4) We do not model packet arrival process at the base station. We assume that the mobiles have infinite backlog at the base station (down link scenario). Such an assumption is standard and is also used in [23, 15, 18].

(P5) We model only the downlink scenario, *i.e.*, traffic flows only from the base station to the mobiles. The extension to handle uplink resource allocation is along similar lines.

4. PF SCHEDULER FOR GENERIC OFDMA BASED RELAY NETWORKS

We start by describing the scheduling framework for general OFDMA based relay networks. The goal of the scheduling algorithm is to allot sub channel-time slot pairs to the relays and mobiles (under suitable constraints to be described soon) so as to maximize

$$\sum_{m \in \mathcal{M}} \frac{d_m}{R_m},$$

where d_m is the total data transmitted to mobile m in the current scheduling frame, and R_m is the data-rate in bits/frame till the previous frame. As discussed in Section 2, this leads to the proportional fair allocation of bandwidth to the mobiles. Next, we describe the constraints the schedule should satisfy. We use the following notation in the remaining.

$$\mathbf{1}_{(l,c)}(t) = \begin{cases} 1 & \text{if link } l \text{ uses sub-channel } c \text{ in slot } t, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Scheduling constraints:
The schedule should satisfy the following constraints:

1. *Flow conservation and orderliness constraint (FCO):* This constraint is the usual flow constraint with the additional requirement that all the data that a relay node receives in frame is also sent out in the same frame (refer to Property **P3** in Section 3). Specifically, this constraint says that the total data arrived at a relay node u until time slot t must exceed the total data transmitted over its child links L_u till time t.

$$\sum_{t'<t} \sum_{c \in \mathcal{C}} \mathbf{1}_{(l_u,c)}(t') r_{(l_u,c)} \geq \sum_{l \in L_u} \sum_{t' \leq t} \sum_{c \in \mathcal{C}} \mathbf{1}_{(l,c)}(t') r_{(l,c)}, \quad \forall u \in \mathcal{R}, t < N \quad (2)$$

$$\sum_{t' \leq N} \sum_{c \in \mathcal{C}} \mathbf{1}_{(l_u,c)}(t') r_{(l_u,c)} = \sum_{l \in L_u} \sum_{t' \leq N} \sum_{c \in \mathcal{C}} \mathbf{1}_{(l,c)}(t') r_{(l,c)}, \quad \forall u \in \mathcal{R} \quad (3)$$

2. *Transmit-receive constraint (TR):* If a relay has a single transceiver, it cannot transmit and receive concurrently. This constraint requires that a relay node cannot be transmitting on any sub-channel over any of its child links while it is receiving a packet on any sub-channel over its parent link. In our notation, it can be expressed as follows:

$$\max_{c \in \mathcal{C}} \mathbf{1}_{(l_u,c)}(t) + \max_{c \in \mathcal{C}, l \in L_u} \mathbf{1}_{(l,c)}(t) \leq 1 \quad \forall u \in \mathcal{R} \quad (4)$$

3. *Spectrum sharing constraint (SS):* Finally, the spectrum sharing constraint states that, in a given time slot t, a sub-channel can only be used by one link. (refer to Property **P2** in Section 3)

$$\sum_{l \in \mathcal{L}} \mathbf{1}_{(l,c)}(t) \leq 1, \ \forall c \in \mathcal{C} \quad (5)$$

Problem statement:
We are now in a position to state the problem *of proportional-fair scheduling for OFDMA relay networks* (**PSOR**).

Given: A tree topology with the base station as the root and the relay nodes as the intermediate links, the sustainable data rates of each of the links for every sub-channel, and the average data rate R_m that each mobile m has received till the previous scheduling frame.

To find: A complete schedule for the scheduling frame, *i.e.*, variables $\mathbf{1}_{(l,c)}(t)$ subject to constraints in Eq. (2)-(5) such that we maximize the objective function $F(\underline{d})$ given by,

$$F(\underline{d}) = \sum_{m \in \mathcal{M}} \frac{d_m}{R_m}, \quad \text{where } d_m = \sum_t \sum_c r_{(l_m,c)} \mathbf{1}_{(l_m,c)}(t).$$

While the problem looks suspiciously similar to max-flow problems, the FCO and the SS constraint make the problem hard to tackle. In the following, we first discuss the hardness of the problem followed by an LP relaxation and our proposed algorithms.

4.1 Hardness Result

The problem PSOR is not just NP-hard, but we cannot hope to approximate it within $(C/4)^{1-\epsilon}$ (recall that C is the number of sub-channels) for arbitrary relay networks.

THEOREM 1. *For any $\epsilon > 0$, the scheduling problem PSOR cannot be approximated within a factor $(C/4)^{1-\epsilon}$ of the optimal in polynomial time unless problems in NP can be solved in probabilistic polynomial time*[5].

The proof follows from reducing the problem of finding a maximum weight independent set to an instance of PSOR [6]. Theorem 1 shows that, the complex dependence between the sub-channel rates at different links makes the problem difficult. Note that although the typical number of hops in a relay network is no more than 2 or 3, the combinatorial complexity is due to the large number of sub-channels over each hop. The wider the channel bandwidth,

[4]Studying the scheduling problem in relay networks under queueing model is part of our future work.

[5]Problems in NP are conjectured and believed to be not solvable in probabilistic polynomial time. The conjecture is open and seems as difficult as the question of whether $P \neq NP$. If we simply assume $P \neq NP$, then PSOR cannot be approximated within $(C/4)^{1/2-\epsilon}$ in polynomial time.

the larger is C. For example, for a 20-50 MHz channel, the number of sub-channels could be as high as 64 or more. In light of the hardness result, we cannot hope to have an algorithm with provable worst case performance bounds for arbitrary relay networks. In the next subsections, we provide algorithm for computing performance upper bound and a heuristic that performs well for most realistic scenarios.

4.2 An Easy to Compute Upper Bound on the Performance

While the PSOR problem is NP-hard, in this subsection, we develop a computationally light upper bound on the performance of any feasible schedule in OFDMA-based relay networks. The upper bound provides an easy-to-compute benchmark against which we compare the scheduling algorithms developed in rest of this section and the next section. Clearly, if we can demonstrate that the performance of any scheduling algorithm is within, say $x\%$ of this upper bound, we can be definite that it is also within $x\%$ of the optimum.

It is not hard to see that, if we solve the LP obtained by dropping the integrality constraints of the variables, the resulting LP solution provides an upper bound on any feasible solution. However, a moment's reflection shows that, such an LP consists of LCN variables and $LN + CN + C^2LN$ constraints. This could give rise to a prohibitively large complexity even to solve the LP with readily available LP-solver tools. In the following, we propose a modified LP with LC variables and $L+C$ constraints, such that the modified LP is guaranteed to produce a solution within 50% of the original LP.

Define $\rho_{(l,c)}$ as, the fraction of time for which link l uses sub-channel c during the given frame.

$$\rho_{(l,c)} = \frac{1}{N}\sum_{t} \mathbf{1}_{(l,c)}(t)$$

LEMMA 4.1. *Any feasible time fraction allocation vector $\rho_{(.,.)}$ must satisfy the following necessary conditions for flow conservation and schedulability.*

$$\sum_{c\in\mathcal{C}} \rho_{(l_u,c)} \cdot r_{(l_u,c)} \geq \sum_{l\in L_u}\sum_{c\in\mathcal{C}} \rho_{(l,c)} \cdot r_{(l,c)}, \qquad \forall u \in \mathcal{R} \quad (6)$$

$$\max_{c\in\mathcal{C}} \rho_{(l_u,c)} + \max_{c\in\mathcal{C}, l\in L_u} \rho_{(l,c)} \leq 1, \quad \forall u \in \mathcal{R} \quad (7)$$

$$\sum_{l\in\mathcal{L}} \rho_{(l,c)} \leq 1, \qquad \forall c \in \mathcal{C} \quad (8)$$

We skip the details of the proof for want of space. The proof follows by re-writing the constraints in Eq. (2)-(5) by summing up the time slots alloted to every link over each sub-channel. Thus, if we solve the problem of maximizing $F(\underline{d})$ subject to the constraints given by Lemma 4.1, we have an upper bound. However, the constraint given by (7) can be rewritten as linear inequalities using a total of C^2L distinct linear inequalities. Thus, such an LP could still have prohibitive complexity. The following result shows that, if we replace the constraints (7) by

$$\sum_{l\in\mathcal{L}} \rho_{(l,c)} \leq \frac{1}{2}, \qquad \forall c \in \mathcal{C}\ ,$$

then, the resulting solution is guaranteed to be within a factor 0.5 of the solution obtained by LP relaxation of PSOR.

THEOREM 2. *Let C_r be the solution to the following LP:*

Maximize:
$$\sum_{m\in\mathcal{M}} \frac{d_m}{R_m} \quad (9)$$

Subject to:

$$d_m = N \cdot \sum_{c\in\mathcal{C}} r_{(l_m,c)} \cdot \rho_{(l_m,c)}, \qquad \forall m \in \mathcal{M} \quad (10)$$

$$\sum_{c\in\mathcal{C}} \rho_{(l_u,c)} \cdot r_{(l_u,c)} \geq \sum_{l\in L_u}\sum_{c\in\mathcal{C}} \rho_{(l,c)} \cdot r_{(l,c)}, \qquad \forall u \in \mathcal{R} \quad (11)$$

$$\sum_{l\in\mathcal{L}} \rho_{(l,c)} \leq \frac{1}{2}, \qquad \forall c \in \mathcal{C} \quad (12)$$

Let C^* be the solution to the PSOR problem, and let C_{LP} be the solution to the LP corresponding to PSOR. We then have the following:

$$C^* \leq C_{LP} \leq 2C_r$$

PROOF. The proof [6] is based on even-odd scheduling scheme similar to the one proposed in [16], and furthermore requires that the time slot duration be arbitrarily small (fluid flow model). □

REMARK 1. *Finally, we make three important remarks.*

1. The LP given by Theorem 2 has LC variables and $L + C$ constraints, and thus it can be solved very fast using LP solving tools. This provides an easy to compute upper bound on the performance of any scheduling algorithm. We use this as a benchmark to evaluate many of our algorithms.

2. Note that there is no inconsistency between Theorem 1 and Theorem 2 because the LP given in Theorem 2 simply provides an upper bound but does not produce a valid schedule, as the even-odd scheduling algorithm used to prove the bound requires the time slot duration to be arbitrarily small.

3. Theorem 2 gives an upper bound on the objective function of PSOR problem. Another relevant question is, does this also translate into an upper bound on the proportional-fairness metric? Indeed, it can be shown that, if R_m is the average long-run rate to user m under proportional-fair scheduling (R_m's can be had by solving the exact problem of PSOR at every frame), and $R_m^{'}$ is the average long-run rate obtained by solving the LP in the statement of Theorem 2 at every frame, then $\sum_m \ln R_m \leq \sum_m \ln(2R_m^{'})$. We skip the details of this argument.

4.3 A Heuristic ArgMax Algorithm for Relay Scheduling

In this section, we present a simple heuristic that can be viewed as a generalization of the "argmax" based scheduling for networks without relay (*i.e.*, all mobiles connected to a single BS) [5]. *The heuristic consists of two key steps: (i) segmenting the slots in a frame so that the links in the path from BS to mobile lie in different segments, and (ii) assigning sub-channels in different links by giving higher priority to mobiles that use fewer tiles along this path for a unit increment in the $F(\underline{d})$.* Since relay networks are envisioned to be not more than two or three hops, the segmentation of the frame solves the ordering constraint, *i.e.*, the TR constraint in Eq. (4) if we assign all sub-channels for the first hop links to the first segment, all the sub-channels for the second hope links to the second segment, and so on so forth.

We introduce some notations. Let H be the depth of the tree with relays and mobiles. Recall that P_m is the set of links on the path from the base-station to mobile m. Also, we use h_l for the number of hops link l is from the root/base-station. The heuristic is formally described as in Algorithm GenArgMax. The algorithm has three steps:

1. *Segmenting the scheduling frame:* The first part consists of segmenting the frame into H (number of hops) parts (Steps 1-2). The tiles belonging to segment h are only assigned to links that are h hops or more from the BS.

2. *Selecting eligible mobiles:* A mobile m is called *eligible* for scheduling, if for sub-channel c, the mobile has the best ratio of $r_{(l_m,c)}/R_m$ among all the mobiles connected to the relay node (or base-station). This is similar to the argmax scheduling principle [5]. In Steps 3-10 of the Algorithm, we remove all but the eligible mobile for consideration in scheduling.

Algorithm 1 GenArgMax: Generalized ArgMax Scheduling

1: Let H be the maximum hop-count in the network. Let m_h be the number of mobiles using the h^{th} hop, and let $f_h = m_h / \sum_h m_h$. Divide the N slot frame into H segments, where h^{th} segment is of length Nf_h. All sub-channels of link l are scheduled in segment h_l in which l lies. {If Nf_h is not an integer, then we can apply ceiling (floor) for the segments corresponding to even (odd) numbered segments.}
2: Let $s_c(h)$ denote the time-slots available to sub-channel c in slots belonging to segment h. Initialize $s_c(h) := \lfloor Nf_h \rfloor$ for all h.
3: Denote by $\mathcal{C}_h$ the set of available sub-channels in segments h. Initialize $\mathcal{C}_h := \mathcal{C}$ for all h
4: Let $\mathcal{M}_c$ be the mobiles that contend for channels. Initialize $\mathcal{M}_c := \emptyset$.
5: **for all** $u \in \mathcal{R}$ **do**
6: **for all** $c \in \mathcal{C}$ **do**
7: Obtain $m_{(u,c)}$ as the mobile that satisfies

$$m_{(u,c)} = \arg \max_{m \text{ attached to relay } u} \frac{r_{(l_m,c)}}{R_m}$$

8: $\mathcal{M}_c \leftarrow \mathcal{M}_c \cup \{m_{(u,c)}\}$
9: **end for**
10: **end for**{For any relay, we only consider the mobiles that have best r/R over some sub-channel.}
11: **while** ($\mathcal{C}_h \neq \emptyset$ for all h) **do**
12: **for all** $m \in \mathcal{M}_c$ **do**
13: **for all** $l \in P_m$ **do**
14: Compute, $\alpha_m(l)$, the minimum (over available sub-channels) number of slots required by link l for a unit increment in $F(\underline{d})$ due to data transmitted only to mobile m as follows:

$$\alpha_m(l) = \min_{c \in \mathcal{C}_{h_l}} \frac{R_m}{r_{(l,c)}}, \qquad c_m(l) = \arg \min_{c \in \mathcal{C}_{h_l}} \frac{R_m}{r_{(l,c)}}$$

{$c_m(l)$ is the sub-channel used in link l if contribution to $F(\underline{d})$ due to mobile m is incremented in a later step of this iteration}
15: **end for**
16: Compute the maximum slots required by mobile m (in any of the links) for unit increment in $F(\underline{d})$ as follows:

$$\beta_m = \max_{l \in P_m} \alpha_m(l)$$

17: **end for**
18: Let $k = \arg\min_{m \in \mathcal{M}_c} \beta_m$.
19: Find ϵ, the maximum possible increase in $F(\underline{d})$ by transmitting data to mobile k with the available slots. Clearly, ϵ satisfies

$$\epsilon \alpha_k(l) \leq s_{c_k(l)}(h_l), \ \forall\, l \in P_k.$$

Choose $\epsilon = \min_{l \in P_k} s_{c_k}(h_l)/\alpha_k(l)$.
20: **for all** $l \in P_k$ **do**
21: Allocate $\lceil \epsilon \alpha_k(l) \rceil$ slots from segment h_l to link l and sub-channel $c_k(l)$.
22: $s_{c_k(l)}(h_l) \leftarrow s_{c_k(l)}(h_l) - \lceil \epsilon \alpha_k(l) \rceil$
23: **if** ($s_{c_k(l)}(h_l) = 0$) **then**
24: $\mathcal{C}_{h_l} \leftarrow \mathcal{C}_{h_l} \setminus c_k(l)$
25: **end if**
26: **end for**
27: **end while**

3. *Computing most eligible mobile:* This step is repeated till we do not have any sub-channels remaining in any of the segments. In the following, we describe one iteration of this procedure (one iteration of the *while* loop in Steps 11-27). For every eligible mobile, we first obtain the tiles required to increase $F(\underline{d})$ by transmitting data to m only. Now, incrementing $F(\underline{d})$ by one unit by only increasing d_m would require R_m units of data to be transfered to mobile m, which would further require $R_m/r_{(l,c)}$ time-slots on any link l that is in the path to mobile m if sub-channel c alone is used for transmitting this data. Thus, on link l, mobile m needs at least $\alpha_m(l) = \min_c R_m/r_{(l,c)}$ number of slots if the best available sub-channel is used. Thus, the maximum number of slots required on any of the segments is $\beta_m = \max_{l \in P_m} \alpha_m(l)$ as the links in the path lie on distinct segments. We view β_m as the resource required for unit increase in $F(\underline{d})$ by increasing d_m alone. This computation is done in Steps 12-17 of the Algorithm. In every iteration, the mobile m with minimum β_m is chosen (Step 18). Say this mobile is k. In Steps 19-26, based on the available remaining tiles in different segments, we increment $F(\underline{d})$ as much as possible by increasing d_k alone.

4. *Repetition of the steps:* The previous step is repeated till there is some tile available in all the segments.

The following can be shown easily. We skip the details.

PROPOSITION 4.1. *Alg. GenArgMax has complexity* $O(LC)$.

5. IEEE 802.16J SCHEDULING FRAMEWORK

Although the problem formulation in Section 4 aims at finding the optimum resource allocation, it does not take into account the overheads involved in realistic systems. For example, when a radio makes a transition from receive mode to transmit mode, or vice-versa, it needs a non-zero amount of transition time for its power amplifiers to ramp up, and its circuitry to synchronize. Thus, to reduce the system bandwidth loss due to this transition overhead, it is desirable to have a node finish all its transmissions in one transmit opportunity. Motivated by this fact, and furthermore, to simplify scheduling, the IEEE 802.16j working committee has proposed a scheduling framework in the IEEE 802.16j draft [10] in which only one node is allowed to transmit at any given time instant during the downlink subframe[6]. This considerably simplifies the problem formulation as compared to the problem formulation in Section 4 where the problem amounts to solving a two-dimensional packing. The IEEE 802.16j standard makes provision for disseminating the scheduling information during the transmission of the preamble from the base station. By restricting the transmission opportunity to one node at a time, the schedule can be disseminated using lower messaging overheads. Despite this, note that in order to exploit frequency selectivity to the maximum, the optimum schedule may result in significantly different modulation and coding schemes over each hop and across different subchannels. As a result, the overheads of schedule dissemination may become significant. Although the impact of schedule dissemination overheads is not addressed in this work, it is part of our future studies.

In the following, we state the IEEE 802.16j based PSOR problem (called 16jPSOR) problem using the simplification stated above.

Problem statement:
The problem essentially amounts to solving PSOR with the restriction that only one relay (or the BS) can transmit in a time-slot, *i.e.*, the same node transmits over all the sub-channels in any slot. We refer to this problem as **16jPSOR**. This can be formulated as an integer linear program.

Define d_m to be the amount of data sent over to mobile m during the current scheduling frame. Let $n^u_{(l,c)}$ represent the number of slots assigned to transmissions over link l outgoing from node u using sub-channel c. Let t_u be the total number of time slots assigned to node u. The Proportional Fair Scheduling problem in Section 4 can be reformulated within the IEEE 802.16j framework as the follows:

$$\text{Maximize:} \quad F(\underline{d}) = \sum_{m \in \mathcal{M}} \frac{d_m}{R_m} \tag{13}$$

Subject to:

$$\sum_{c \in \mathcal{C}} n^{p(u)}_{(l_u,c)} \cdot r_{(l_u,c)} \geq \sum_{m \in \mathcal{T}_u} d_m, \qquad \forall u \in \mathcal{R} \cup \mathcal{M} \tag{14}$$

$$\sum_{l \in L_u} n^u_{(l,c)} \leq t_u, \quad \forall c \in \mathcal{C}, \forall u \in \mathcal{R}^+ \tag{15}$$

[6]During the uplink subframe, all the child nodes of a single node are allowed to transmit concurrently (over different time-sub-channel resources).

$$\sum_{u \in \mathcal{R}^+} t_u \leq N \tag{16}$$

$$t_u, n^u_{(l,c)} \in \mathbb{Z}, \qquad \forall u \in \mathcal{R}^+, \forall c \in \mathcal{C}, \forall l \in \mathcal{L}. \tag{17}$$

In the above, the inequalities (14) are the flow constraints that ensure that the amount of data that can be received on the parent link of node u should be greater that the total data received my all mobiles in $\mathcal{T}_u$; the inequalities (15) ensure that the total time allocated to the transmissions over different links outgoing from a relay node u does not exceed the total time-slots allocated to node u; and the inequality (16) ensures that the sum of the time-slots allocated to the different relays do not exceed the total number of slots in a frame. Once the time allocations for all the nodes have been determined, orderliness constraint can be easily satisfied by scheduling the nodes in the routing tree in a breadth-first manner.

5.1 Hardness Result

It is fairly straight-forward to show that the problem is NP-hard.

THEOREM 3. *The 16jPSOR is NP-hard even when the channel-gains are sub-channel independent.*

The proof follows by reducing the knapsack problem to an instance of 16jPSOR, and has been omitted due to space constraints.

In the following, we develop two heuristics and in a later section we demonstrate that the heuristics perform very close to the optimal in practical scenarios. Note that the upper bound from Subsection 4.2 also serves as an upper bound to the heuristic algorithms proposed in this section, since the scheduling framework of IEEE 802.16j is a special case of the generic OFDMA-based relay framework considered in Section 4.

5.2 Fast Heuristic PF Scheduling

We now propose a fast heuristic to solve 16jPSOR. The heuristic has two simple steps: *(i) solving the LP corresponding to 16jPSOR under two realistic and simplifying assumptions to be stated soon, followed by (ii) rounding the LP-based solution without violating the constraints of the 16jPSOR problem*. As we will show, the two simplifying assumptions lead to an LP that could be solved in closed form without using any LP-solver tool. The two assumptions are as follows:

1. For any relay node u, transmissions over sub-channel c to mobiles associated with u happens only to the mobile m for which $\frac{r_{(l_m,c)}}{R_m}$ is maximum.
2. We assume that each relay link has a time duration associated to it during which all the sub-channels are used for transmission over that link. In other words, the time allocated to each relay node/BS u is partitioned into multiple segments:
   - $T_m(u)$, the number of time-slots for transmitting data to the mobiles attached to u, and
   - $T_r(l)$, the number of time-slots for transmitting data on the relay-child link $l \in L^{'}_u$.

The first simplification can be proved formally for the mobiles connected to the base-station, and hence this is an extension of this base-station's property to the relay nodes. Thus each mobile receives data exclusively over certain sub-channels. *Let $\mathcal{C}_m$ be the set of sub-channels for which mobile m is eligible*. Among the mobiles associated to relay u, let $m_c(u)$ be the mobile which is assigned sub-channel c.

$$m_c(u) = \text{argmax}_{m \in S_u} \left(\frac{r_{(l_m,c)}}{R_m} \right) \tag{18}$$

The intuition behind the second simplification is that, typically the relay links have clear line of sight and smaller path loss exponent as compared to the relay to mobile or BS to mobile links. Furthermore, since the BS-relay links and the relay-relay links are ART-ART (Above Rooftop to Above Rooftop), the delay spread of the channel is low [8]. This results in **(i)** high SINR for all the sub-channels, and **(ii)** little or no frequency diversity as a result of low delay spread. Consequently, treating the relay links as fat high-speed pipes, and isolating their resource allocation from the resource allocation of mobile links is a reasonable. Accordingly, we define $\mathsf{C}_u, \bar{\mathsf{C}}_u, \mathsf{U}_u$ as follows:

$$\mathsf{C}_u \triangleq \sum_{c \in \mathcal{C}} r_{(l_u,c)}, \;\; \bar{\mathsf{C}}_u \triangleq \sum_{c \in \mathcal{C}} r_{(l_{m_c(u)},c)}, \;\; \mathsf{U}_u \triangleq \sum_{c \in \mathcal{C}} \frac{r_{(l_{m_c(u)},c)}}{R_{m_c}} \tag{19}$$

C_u is the total rate of a relay node u to its parent $p(u)$ (*i.e.*, the rate summed over all the sub-channels), $\bar{\mathsf{C}}_u$ is the *effective capacity* of a relay station/BS to its mobiles, and U_u is the contribution (per-time slot) to $F(\underline{d})$ by the mobiles associated to relay node u. Note that, for a given problem instance (*i.e.*, in a given scheduling frame), $\mathsf{C}_u, \bar{\mathsf{C}}_u, \mathsf{U}_u$ are constants that are completely determined by the rates of all the links over all the sub-channels. We now show how the assumptions simplify the problem using these constants.

FACT 5.1. *Under the two simplifying assumptions we make in this section, the objective function of 16jPSOR can be rewritten as*

$$\sum_{u \in \mathcal{R}^+} T_m(u) \cdot \mathsf{U}_u \tag{20}$$

PROOF. First, note that

$$d_m = T_m(p(m)) \cdot \sum_{c \in \mathcal{C}_m} r_{(l_m,c)} \; . \tag{21}$$

Recall that S_u is the set of mobiles directly attached to node u. $F(\underline{d})$ can be rewritten as follows:

$$\begin{aligned}
\sum_{u \in \mathcal{R}^+} \sum_{m \in S_u} \frac{d_m}{R_m} &= \sum_{u \in \mathcal{R}^+} \sum_{m \in S_u} \frac{T_m(p(m))}{R_m} \cdot \sum_{c \in \mathcal{C}_m} r_{(l_m,c)} \\
&= \sum_{u \in \mathcal{R}^+} T_m(u) \sum_{m \in S_u} \sum_{c \in \mathcal{C}_m} \frac{r_{(l_m,c)}}{R_m} \\
&= \sum_{u \in \mathcal{R}^+} T_m(u) \sum_{c \in \mathcal{C}} \frac{r_{(l_{m_c(u)},c)}}{R_{m_c(u)}} \\
&= \sum_{u \in \mathcal{R}^+} T_m(u) \cdot \mathsf{U}_u.
\end{aligned}$$

□

It is not hard to see that, under the simplifying assumptions, the constraints of the 16jPSOR problem can be expressed in terms of problem constants $\mathsf{C}_u, \bar{\mathsf{C}}_u, \mathsf{U}_u$'s and the variables $T_m(u)$'s and $T_r(l)$'s. Since the total amount of data transmitted by relay node/BS u to the mobiles directly attached to node u is $T_m(u) \cdot \bar{\mathsf{C}}_u$, the flow constraint in Eq. (14) can now be rewritten as follows:

$$\sum_{v \in \mathcal{T}_u \cap \mathcal{R}} T_m(v) \cdot \bar{\mathsf{C}}_v \leq T_r(l_u) \cdot \mathsf{C}_u \qquad \forall u \in \mathcal{R} \tag{22}$$

Also, since the time allocation of a relay station/BS is partitioned into $T_m(u)$ and $T_r(l)$, the constraints in Eq. (15)-(16) reduce to the following.

$$\sum_{u \in \mathcal{R}^+} T_m(u) + \sum_{u \in \mathcal{R}} T_r(l_u) \leq N \tag{23}$$

Thus, with the two simplifying assumptions, the 16jPSOR problem reduces to maximizing the expression (20) subject to the constraints (22) and (23). We also have the additional constraints that $T_m(u)$'s and $T_r(l)$'s are integers to ensure that the time allocations are integral multiples of a time slot duration. Note that, we have reduced the original problem to one with $2R+1$ variables and $R+1$

constraints, which is independent of the number of sub-channel C. While Mixed Integer LP solvers can be used to solve the problem, we take advantage of the simplicity of the LP relaxation of this problem. To, see this, first note that, under LP relaxation, inequality (22) reduces to an equality[7], in which case (22) and (23) can be combined to form a single inequality. We summarize this observation in the form of the following:

FACT 5.2. *Under the two simplifying assumptions we make in this section, the LP relaxation of 16jPSOR can be expressed as follows:*

$$\textit{Maximize} \sum_{u\in\mathcal{R}^+} T_m(u)\cdot \mathsf{U}_u \tag{24}$$

$$\textit{Subject to:} \sum_{u\in\mathcal{R}^+} \left(\frac{\bar{\mathsf{C}}_u}{\hat{\mathsf{C}}_u}\right) T_m(u) \le N\,, \tag{25}$$

where,

$$\hat{\mathsf{C}}_u = \left\{\frac{1}{\bar{\mathsf{C}}_u} + \sum_{v\in P'_u}\frac{1}{\mathsf{C}_v}\right\}^{-1} \qquad \forall u\in\mathcal{R}^+\,, \tag{26}$$

and P'_u is the set of relay nodes that belong to the path from relay node u to the BS. We assume that $u\in P'_u$, and BS $\notin P'_u$.

PROOF. Under LP relaxation, the inequality (22) reduces to an equality, in which case (22) and (23) can be combined to form a single inequality as follows:

$$\sum_{u\in\mathcal{R}^+} T_m(u) + \sum_{u\in\mathcal{R}}\sum_{v\in\mathcal{T}_u}\left(\frac{\bar{\mathsf{C}}_v}{\mathsf{C}_u}\right)\cdot T_m(v) \le N \tag{27}$$

After additional rearrangement of terms, the preceding inequality can be rewritten to arrive at the stated result. □

The solution to the LP relaxation in Fact 5.2 can be easily obtained in one step as follows:

$$T_m(u) = \begin{cases} \frac{\hat{\mathsf{C}}_u}{\bar{\mathsf{C}}_u}\cdot N & \text{if } u = \text{argmax}_{u\in\mathcal{R}^+}\left\{\frac{\mathsf{U}_u\cdot\hat{\mathsf{C}}_u}{\bar{\mathsf{C}}_u}\right\},\\ 0 & \text{otherwise.}\end{cases} \tag{28}$$

Thus, we observe that the optimum solution consists of serving all the mobiles attached to one relay station/BS, and this choice is determined by the quantity $\mathsf{U}_u\hat{\mathsf{C}}_u/\bar{\mathsf{C}}_u$. The slots alloted to an intermediate link l_v on the path between the BS and the optimum node u is simply $N\hat{\mathsf{C}}_u/\mathsf{C}_v$.

We formally state the heuristic algorithm in Algorithm FastHeuristic16j where we also show how the LP rounding can be done without violating the flow constraints. Step 1-Step 3 compute the different constants and Step 4 determines the optimum relay station whose mobiles are to be served. In Step 5, while performing time allocation using the LP-solution in (28), we use $N-H$ slots instead of N slots. This ensures that we have H spare slots to round off the time-allocations of the relays to *ceiling*. We also round off the time-allocation of the mobiles to *floor*, and so we do not violate the flow constraints (because the capacity of relay links are improved after rounding, while the capacity of mobile links is reduced after rounding). The following result is immediate from Algorithm FastHeuristic16j.

PROPOSITION 5.1. *Algorithm FastHeuristic16j has complexity $O(LC)$.*

[7]This is because, if for a certain link l_u, Eq. (22) is a strict inequality, then the difference between the RHS and the LHS can be deducted from $T_r(l_u)$, and this time can be used to transmit data to a mobile, thereby increasing the value of the objective function.

Algorithm 2 FastHeuristic16j: Fast Heuristic PF Scheduling

1: For $u\in\mathcal{R}^+, c\in\mathcal{C}$, determine the eligible mobile $m_c(u)$ using Eq. (18).
2: For each $m\in\mathcal{M}$, determine the set of sub-channels, $\mathcal{C}_m$, the mobile m is eligible for.
3: For $u\in\mathcal{R}$, determine C_u as defined in (19). For $u\in\mathcal{R}^+$, determine $\mathsf{U}_u, \bar{\mathsf{C}}_u, \hat{\mathsf{C}}_u$ as defined in (19), and (26) respectively.
4: Determine the relay/BS u^* whose mobiles are to be served as follows.

$$u^* = \text{argmax}_{u\in\mathcal{R}^+}\left\{\frac{\mathsf{U}_u\cdot\hat{\mathsf{C}}_u}{\bar{\mathsf{C}}_u}\right\}$$

5: For all $v\in P_{u^*}$, do the following rounding:

$$H := \{\text{Hop count from BS to } u^*\}$$

$$T_m(u^*) \leftarrow \left\lfloor \frac{\hat{\mathsf{C}}_{u^*}}{\bar{\mathsf{C}}_{u^*}}\cdot(N-H)\right\rfloor\,,\; T_r(l_v) \leftarrow \left\lceil \frac{\hat{\mathsf{C}}_{u^*}}{\mathsf{C}_v}\cdot(N-H)\right\rceil$$

$$\Delta := N - \left(T_m(u^*) + \sum_{v\in P_{u^*}} T_r(l_v)\right)$$

6: **if** $\Delta > 0$ **then**
7: Allocate the leftover slots to mobiles attached to the BS.

$$T_m(0) \leftarrow T_m(0) + \Delta$$

8: **end if**
9: Determine the amount of data to be transmitted to each mobile

$$d_m = \sum_{c\in\mathcal{C}_m} T_m(p(m))\cdot r_{(l_m,c)} \qquad \forall i\in S_{u^*}\cup S_0$$

5.3 Heuristic Tree-traversal Scheduler

In this section we describe another simple to implement heuristic scheduler. FastHeuristic16j is very simple to implement, but it is suitable when there is little or no frequency selectivity for the relay links. The algorithm proposed in this subsection is suitable when even the relay links have frequency selectivity, but it has a slightly higher running time compared to FastHeuristic16j as we show later in our results. The heuristic solves the optimal allocation problem under the assumption that, *a sub-channel at a relay node, say u, is dedicated to transmission to only one of the child nodes (which could be a mobile associated to u or another relay node) of u.* Given this assumption, the algorithm works by, traversing the routing tree in a bottom up manner and computing for every RS/BS u, the fraction of time-slots to be assigned to RS's in $\mathcal{T}_v$ (for all relays v that is a child of u) out of every unit time-slot allocated to $\mathcal{T}_u$. For every node $u\in\mathcal{R}^+$, the heuristic computes three quantities i_u, c_u, and t_u as defined below.

i_u = For every unit time-slot allocated to the entire subtree $\mathcal{T}_u$, the increase in $F(\underline{d})$ due to data transmitted to mobiles in $\mathcal{T}_u$.

c_u = The total data per time-slot that has to be transmitted to the mobiles in subtree $\mathcal{T}_u$ for incrementing $F(\underline{d})$ by i_u per time-slot.

t_u = Fraction of transmission time allocated to relays in $\mathcal{T}_u$ for every unit time allocated to$\mathcal{T}_{p_u}$.

We later show how i_u and c_u can be used to obtain the fraction of time $\mathcal{T}_u$ gets from the total time-allocation to $\mathcal{T}_{p_u}$. The algorithm works by traversing the routing tree in a bottom-up manner by first working on the leaf-nodes of the tree. The algorithm is formally described in Algorithm 3. It can be described in three steps as below:

1. *Computing i_u and c_u for the leaf relay-nodes:* (Step 1-Step 6

Algorithm 3 TreeTraversingScheduler: A Tree-traversing scheduler for 802.16j

1: **for all** $u \in \mathcal{R}$ that is a leaf **do**
2: **for all** $c \in \mathcal{C}$ **do**
3: Find the mobile m associated to u for which $R_m/r_{(l_m,c)}$ is least. Let this mobile be $n_c(u)$ which receives data from u over sub-channel c.
4: **end for**
5: **end for**
6: Compute c_u and i_u according to (29).
7: **for all** Intermediate relay node u reached by traversing the tree in a bottom-up manner **do**
8: **for all** $c \in \mathcal{C}$ **do**
9: For every relay v such that $p_v = u$, compute t_v the slots required for unit unit increment in cost due to the mobiles in $\mathcal{T}_v$ as follows:

$$t_v = \frac{c_v}{(i_v r_{(l_v,c)})} + \frac{C}{i_v} .$$

10: For every mobile m associated to u compute $t_m = R_m/r_{(l_m,c)}$.
11: Find $n_c(u) = \arg\min_{v,m}\{t_v, t_m\}$. {Relay u only transmits to $n_c(u)$ on sub-channel c.}
12: **end for**
13: Find the time-fraction $t_u{}'$ allocated for transmission by relay u out of time allocated for the transmission of nodes in $\mathcal{T}_u$ as

$$t_u{}' = \frac{1}{1 + \sum_{v:p_v=u} \frac{r_v}{c_v}}.$$

14: For every relay v such that $p_v = u$, obtain the data/time-slot from u as

$$r_v = \sum_c r_{(l_v,c)} \mathbf{1}_{\{n_c(u)=v\}}$$

and the fraction of transmission time allocated to $\mathcal{T}_v$ out of the time allocation to $\mathcal{T}_u$ as

$$t_v = t_u{}' \frac{r_v}{c_v} .$$

15: Compute c_u and i_u according to the expressions given in (30) and (31).
16: **end for**
17: Traverse the tree in a top-down manner and allocate exact time-slots for which each relay transmits. For example, while doing the allocation for relay u, if we have already allocated S_u slots to $\mathcal{T}_u$, then allocate $\lfloor S_u t_v \rfloor$ time-slots to $\mathcal{T}_v$ for every $p_v = u$. Allocate $S_u - \sum_{v:p_v=u} \lfloor S_u t_v \rfloor$ time-slots for the transmission of u.

of Algorithm 3) For every leaf relay-node (*i.e.*, relay nodes whose all the children are mobiles and none are relays), the algorithm allocates every sub-channel to the mobile that requires least number of tiles for unit increment in $F(\underline{d})$. Clearly, for c^{th} sub-channel, the number of tiles required by mobile m for unit increment in $F(\underline{d})$ is $R_m/r_{(l_m,c)}$. For each sub-channel the heuristic picks the mobile m for which $t_m = R_m/r_{(l_m,c)}$ is least. Let m_c be the mobile selected for sub-channel s_c. Then, c_u and i_u can be obtained as follows:

$$c_u = \sum_c r_{(l_{m_c},c)} \ , \ i_u = \sum_c r_{(l_{m_c},c)}/R_{m_c} \tag{29}$$

2. *Computing i_u and c_u for the intermediate (non-leaf) relay nodes:* (Step 7-Step 16) For an intermediate relay node u, for every sub-channel, we find the node (among the relays and the mobiles associated to u) to which u transmits. Consider a relay node v whose parent is u. We find the number of tiles used by the nodes in $\mathcal{T}_u$ for unit increment in $F(\underline{d})$ if u transmits only to v using c^{th} sub-channel alone. Clearly, from the definition of c_v and i_v, c_v/i_v is the amount of data required to be transmitted to v for unit increment in $F(\underline{d})$. To transmit this amount of data over sub-channel c, relay u needs to transmit to v over sub-channel c for $c_v/(i_v r_{(l_v,c)})$ slots (tiles). In addition, relay nodes in $\mathcal{T}_v$ also require a time-allocation of $1/i_v$ time-slots, or equivalently C/i_v tiles, for unit increment in cost. Thus, the number of tiles used by the nodes in $\mathcal{T}_u$ for unit increment in the objective function if u transmits to v over sub-channel c is $t_v = c_v/(i_v r_{(l_v,c)}) + C/i_v$. Also, for mobile m associated to u, $t_m = R_m/r_{(l_m,c)}$ tiles are required for unit increment in $F(\underline{d})$. Relay u transmits to the node for which t_m (corresponding to the mobiles) or t_v (corresponding to the child relays of u) is least. We call this node $n_c(u)$ (or simply n_c when there is no ambiguity) which could be a mobile or a relay.

We now wish to find the values of c_u and i_u. For every v that is a child of u, first obtain the data rate (denoted by r_v) at which u transmits to v. Clearly,

$$r_v = \sum_c r_{(l_v,c)} \mathbf{1}_{\{n_c(u)=v\}} .$$

Let t_v be the fraction of time allocated to $\mathcal{T}_v$ for every unit time-slot allocated to relay nodes in $\mathcal{T}_u$. Also, let $t_u^{'}$ be the fraction of time for which u transmits for every unit time allocated to relay nodes in $\mathcal{T}_u$. Clearly, by conservation of flows, $r_v t_u{}' = c_v t_v$. Also since $t_u^{'} + \sum_v t_v = 1$, we have

$$t_u^{'} = \frac{1}{1 + \sum_v \frac{r_v}{c_v}}, \ t_v = t_u{}' \frac{r_v}{c_v} .$$

Also,

$$c_u = t_u{}' \sum_c r_{(l_{n_c},c)} \ , \tag{30}$$

$$i_u = t_u{}' \left(\sum_{c:n_c \in \mathcal{R}} \frac{r_{(l_{n_c},c)} i_{n_c}}{c_{n_c}} + \sum_{c:n_c \in \mathcal{M}} \frac{r_{(l_{n_c},c)}}{R_{n_c}} \right) . \tag{31}$$

The $t_u^{'}$ factor appears in the preceding because node u actually transmits for a $t_u^{'}$ time out of every unit time allocated to $\mathcal{T}_u$.

3. *Computing the time-allocations:* (Step 17) Note that, upon traversing the entire tree in a bottom-up manner, we have already computed t_u's for all the nodes. Since we know that all the N slots are available to the tree rooted at the base-station, we can now traverse the tree in a top-down manner and calculate the exact time-allocations for every relay.

The following result is fairly straightforward.

PROPOSITION 5.2. *Algorithm TreeTraversingScheduler has complexity $O(LC)$.*

6. PERFORMANCE EVALUATION

In this section, we present simulation results to evaluate the performance of the proposed scheduling algorithms. The goal of this section is three-fold. First, to quantify how much off our scheduling algorithms are from the optimal. Second, to compare our approach with other approaches proposed in literature, and third, to understand the benefits of relays in enhancing throughput, and extending range.

6.1 Simulator Setup

In this subsection, we give a brief overview of the different modules used in our custom simulator.

Network topology:
We simulate a single 120 degree sector in a WiMAX cell. Relay node locations are judiciously chosen so as to provide a uniform coverage in the cell-sector. The exact relay-locations are different for different settings, and are provided along with the results. Except for the results on range extension, the default cell-radius is 1 km which is typical coverage of WiMAX base stations for urban environment [22]. The mobile stations (MS) are distributed randomly and uniformly over the cell-sector. Multi-hop shortest path

routes to the BS are determined by using the Expected Transmission Time or the ETT routing metric [19].

Subchannelization:
We assume a bandwidth of 10 MHz at carrier frequency of 2.5 GHz. We assume a 1:3 spatial reuse in which the 10MHz spectrum is split into three segments, and each segment is used in a single sector. Details about the number of sub-carriers and sub-channels can be found in Table 2. We assume Band AMC (Adaptive Modulation and Coding) operation [9] in which a sub-channel is formed from 54 contiguous sub-carriers of which 48 sub-carriers are used for sending data, and the rest are for pilot signal.

Path-loss and Frequency-selective fading:
For modeling the wireless channel, we have incorporated several proposals by IEEE 802.16j task force [8, 10]. For the BS-RS and the RS-RS links, we use the Type H line-of-sight path loss model which is recommended for ART-ART (above rooftop to above rooftop) urban links, while for the BS-MS and the RS-MS links, we use Type E non-line-of-sight path loss model which is recommended for ART-BRT (above rooftop to below rooftop) urban links [8]. Parameters for path-loss and log-normal shadowing are chosen as per simulation evaluation methodology recommended in [8, 10]. In order to simulate frequency selective fading, we take the following approach. The extent of frequency selectivity is determined by the delay spread of the channel which is a measure of the time duration over which most of the replicas of the transmitted signal reach the receiver. The higher the delay spread, the more the extent of frequency selective fading of the channel [20]. For the ART-ART urban links, a delay spread of 0.111 μs, and for the ART-BRT urban links, a delay spread of 1.257 μs has been reported via extensive measurements [8]. We determine the 50% coherent bandwidth of the channel, *i.e.*, the bandwidth, B_c over which the fading channel gains have a correlation of less than 50% [20]. We then partition the entire 10MHz bandwidth into blocks of size B_c, and assume that the channel gain is constant over each such block, and is identically and independently distributed over different blocks. Three independent Rayleigh/Ricean waveforms are generated for each block, and a weighted sum of these signals is taken to implement the multi-tap fading model. The fading waveforms are generated using the modified Jakes fading model. Once the SINR of each component block of a sub-channel is determined, we use the SINR to rate mapping from [9] for the modulation coding schemes supported under IEEE 802.16e to determine the rate of each block. The rates of component blocks are then added up to determine the rate of a sub-channel.

Miscellaneous settings:
We assume pedestrian users with a speed of 3.0 kmph. This also models the fading channel of a static user because of the non-static environment. We do not present results for vehicular users (speed of 60 to 100 kmph). The channel gains of vehicular users change significantly faster as compared to the frame duration, and therefore are not amenable to exploiting frequency selectivity. A fixed fraction of slots can be reserved for vehicular users (using diversity permutation mode, see Section 2), and the scheduling problem of vehicular and pedestrian users can be easily decoupled. Therefore, we only focus on the scheduling of pedestrian users. Velocity-dependent time correlation between shadowing gains of a mobile across time is determined using Gudmundson's model. Important simulation parameters are listed in Table 2.

Evaluated algorithms :
We evaluate the following algorithms: (i) GenArgMax, (ii) FastHeuristic16j, (iii) Tree-traversing, (iv) Round-robin, and (v) OFDM^{2}A. In Round-robin scheduling, one MS is chosen, and is served for the entire frame duration. OFDM^{2}A is a scheduling algorithm proposed for relay networks [18]. OFDM^{2}A tries to exploit the frequency diversity of the sub-channels over different links. However, as we discussed in detail in Section 1, it does not satisfy the Transmit-Receive constraint, *i.e.*, under OFDM^{2}A, a node may be transmitting and receiving on different sub-channels at the same time. Nevertheless, we include OFDM^{2}A for the sake of comparison. For each topology in each frame, we also solve the LP in Theorem 2 which we refer to as half-approximate LP. The throughput obtained using the half-approximate LP is multiplied by two to obtain an upper bound on the optimum.

Parameter	Value
BS transmit power	43dBm (20 watts)
RS transmit power for throughput-enhancement	37dBm (5 watts)
RS transmit power for range extension	40dBm (10 watts)
BS-RS, RS-RS shadowing standard deviation	3.5dB
BS-MS, RS-MS shadowing standard deviation	8dB
BS, RS antenna gain	15dB
Noise power	-174 dBm/Hz
BS height	30 m
RS height	15 m
MS height	1.5 m
Number of MS	40
Sub-carrier bandwidth	10.94 KHz
Sub-carriers per sub-channel	54
Number of sub-channels per sector	5
Frame Duration	5ms
Slots per frame	48
Simulation duration	10 s (2000 frames)

Table 2: Simulation parameter settings.

6.2 Simulation Results

6.2.1 Near-optimality of heuristic algorithms

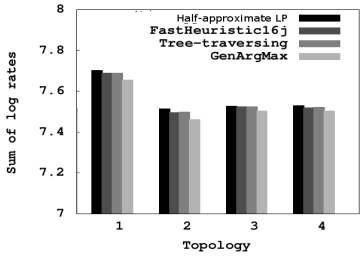


Figure 2: Proposed algorithms perform close to the optimum which is upper bounded by the twice the half-approximate LP in Theorem 2.

Scheduling algorithm	Average running time (ms)
FastHeuristic16j	0.013
Tree traversal	0.026
GenArgMax	0.034
OFDM^{2}A	0.025
Round Robin	0.002
LP-based bound in Theorem 2	18

Table 3: Proposed algorithms take less than 0.05 ms during each scheduling run.

First we consider the setting for which the sector radius is 1 km, and 3 relays are placed equidistant along an arc of radius 0.8 km. We plot the objective function (sum of the log of the long term throughput of all the mobiles) for the proposed scheduling algorithms in Fig. 2 for representative topologies. Note that a proportional fair scheduler optimizes this quantity as discussed in Subsection 2.2. We summarize our observations as follows.

- Fig. 2 shows that all the proposed heuristic algorithms perform close to the optimum (off by less than 0.5%). Extensive simulations for more random topologies show identical performance, and have been omitted due to space constraints.
- Table 3 shows that the average running time of the proposed heuristic algorithms are of the order of microseconds (over Intel Centrino Core 2 Duo machine running at 2GHz, a memory of

1GB, and without any multi-threading), and hence the scheduling deadline of 5 ms can be easily met. The running times were determined by timing the scheduler component of the simulator after all the channel information has been collected. We also note that the computation time of LP based upper bound in Theorem 2 is 18 ms using the open source GNU Linear Programming Kit [1]. While faster LP solvers are available, this points out that schedulers based on rounding LP solutions may not be feasible for 5 ms frame duration of WiMAX. Furthermore, solving LP problems with integrality constraints (the original scheduling problem) takes even longer time due to the inherent hardness of the problem.

- The restriction imposed by IEEE 802.16j framework to allow only one active node at a time does not result in significant performance degradation, since FastHeuristic16j and Tree-traversing algorithms (which adhere to IEEE 802.16j standard) perform close to optimal.

6.2.2 Throughput enhancement

In this subsection, we consider FastHeuristic16j as our representative scheduling algorithm, and compare its performance (median and mean throughput) with other scheduling algorithms proposed in the past, as well as with the no-relay scenario. For the no-relay scenario, we implement the following scheduling policy that leads to PF allocation of rates: over such sub-channel c, BS transmits to the mobile m for which $r_{(l_m,c)}/R_m$ is maximum. We consider four representative random topologies, and plot the CDF of mobile throughput in Fig. 3(a)-3(d), and the median and the mean of mobile throughput in Fig. 4(a)-4(b). We note the following.

- Fig. 3(a)-3(d) show that when relays are employed, both the mean (area between the curve and the y-axis), as well as the median improve.
- The median throughput improves by about 20-30% for all the four random topologies, while the mean throughput improves by about 10-20% over the no-relay scenario.
- As discussed in Section 1, OFDM^{2}A requires multiple transceivers. Despite this, we note from Fig. 4(a)-4(b) that our proposed algorithm, FastHeuristic16j either performs as well as, or better than OFDM^{2}A in most cases.
- Even with relays, Round Robin performs the worst as it does not exploit multi-user diversity and frequency selectivity.

We also run additional simulation runs for 36 randomly generated topologies, *i.e.*, 36 different randomly generated mobile locations. In Fig. 4(c)-4(d), we plot the percentage improvement in the median and the mean throughput over no-relay scenario when FastHeuristic16j (a representative of our proposed algorithms) is used. We note the following.

- Fig. 4(c) shows that that for more than 50% cases, the median improves by at least 15%, and by as much as 35%. For more than 78% cases, the median improves by at least 5%.
- There is a single scenario for which the median drops (by less than 5%). We studied this run closely, and found that this behavior was due to the frequency of association updates for mobiles which was set to be once every 10 frames. A higher update frequency results in optimal associations, but has higher signaling overheads.
- Fig. 4(d) shows that for more than 30% cases, the mean improves by at least 10%, and by as much as 15%. Furthermore, in more than 82% cases, the mean improves by at least 5%.

Similar results were observed for GenArgMax and Tree-traversing, and have been omitted due to space constraints. *The above results show that although relaying does not result in significant benefits for every random topology, there are several topologies where the mobiles have poor shadowing and path loss gains, and the improvements for these scenarios with relays are significant. These observations justify the deployment of relays for throughput enhancement by filling coverage holes. Furthermore, our proposed algorithms enable us to exploit the multiuser diversity and frequency selectivity, and provide significantly higher mean and median throughput for these scenarios.*

Sector-radius	Number of Relays	Median Throughput (in kbps)	Mean Throughput (in kbps)	%-Mobiles getting Coverage
1.0 km	3	217	200	100%
1.2 km	5	167	173	95%
1.6 km	7	120	140	90%

Table 4: Coverage-throughput trade-off using relays

6.2.3 Range extension

In this subsection, we present simulation results for the range extension scenario. We run simulations for 36 independent random topologies for the following scenarios: (i) 1km sector radius, 3 relays, (ii) 1.2km sector radius, 5 relays, and (iii) 1.6km sector radius, 7 relays. When the sector radius is 1.2km, two relays are placed at 0.8km, and the rest three are placed at 1.1km. When the sector radius is 1.6km, two relays are placed at 0.8km, two relays at 1.1km, and the rest three relays at 1.4km. This relay deployment is used to provide uniform coverage across the entire sector. We do not study the problem of optimum relay placement, since relay placement depends on results of local site survey (location of obstacles, etc.), and is part of our future work. Table 4 shows the throughput and coverage vary with the radius. We note the following.

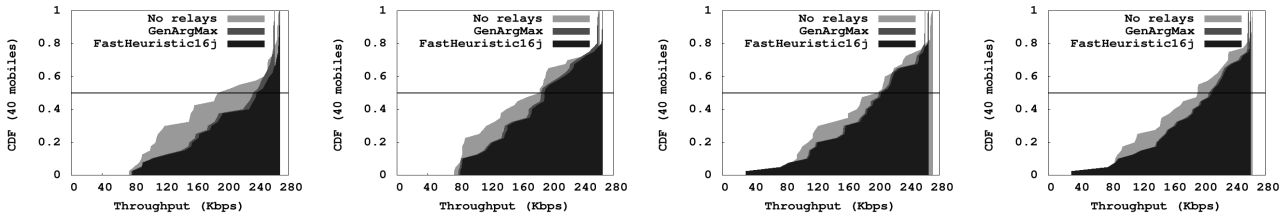
- Range extension to 1.2km is possible with 5 relays at the cost of a slightly lower median throughput of 167kbps, and 95% coverage can be guaranteed for this scenario, *i.e.*, less than 5% of the mobiles cannot be served due to poor channel conditions.
- Range extension to 1.6km is possible with 90% coverage, and a median throughput of 120kbps can be guaranteed.

The above results show that while 100% coverage may not be possible with as many as 5-7 relays, the use of relays provides an attractive incremental *solution to expanding an operator's network. To begin with, the distant locations can be reached using relays. However, as the demands of the remote location increase over a period of time, a dedicated base station can be installed in that location. Thus, using relays, the network expansion can be carried out in a more cost-efficient and incremental manner.*

7. RELATED WORK

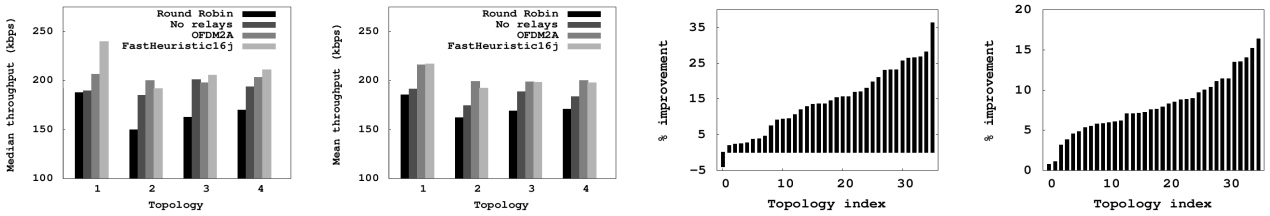
The IEEE 802.16j task group is currently working on the standardization of MAC layer for WiMAX based multi-hop relay networks[10]. In [4], the authors study the problem of scheduling in multi-hop relay networks under a TDMA model. Uplink scheduling that accounts for traffic variations is studied in [11]. However, none of these works include frequency selectivity across sub-channels in their problem formulation. As we have proved in Theorem 1, frequency selectivity across sub-channels substantially increases the complexity of the scheduling problem.

Several works have studied the problem of scheduling, rate and power allocation for OFDM based single hop networks [21, 5, 14, 7]. The authors in [3] consider an OFDMA based single hop system with queueing taken into account (instead of an infinite backlog model). The authors show that the tiling structure of an OFDMA frame results in high combinatorial complexity of the scheduling problem, and then propose approximation algorithms to solve the problem. The work in [12] considers designing OFDMA scheduling frame for a single hop network when the PHY-profiles to be used for the PDUs are known in advance. While these works take into account the frequency selectivity, there is no easy extension of these results to multi-hop relay scenario which is the focus of our work. A family of scheduling disciplines including proportional fair (PF) scheduling is proposed in [23] for opportunistic scheduling in single hop CDMA and TDMA networks where fre-



(a) Topology 1 (b) Topology 2 (c) Topology 3 (d) Topology 4

Figure 3: CDF of mobile throughput for four different random topologies. While relays always result in some improvement in the mean/median throughput, certain scenarios dominated by high path loss and poor shadowing benefit the most from relaying, e.g., topology 1 in Fig. 3(a).



(a) Median Throughput (b) Mean Throughput (c) % improvement in median throughput over no-relay (d) % improvement in mean throughput over no-relay

Figure 4: Relaying results in significant improvement in median and mean throughput, especially in scenarios of high path loss and strong shadowing. Fig. 4(a)-4(b) show relative performance improvement of different scheduling algorithms for four topologies. Fig. 4(c)-4(d) show performance improvement of over the no-relay scenario for additional 36 random topologies demonstrating the benefits of relays for throughput enhancement.

quency selectivity cannot be exploited. For a good survey of PF and other scheduling principles in non-OFDMA networks, we refer the reader to [2] and the references therein.

To the best of our knowledge, [18] is the only work which studies multi-user scheduling for relay networks in the presence of frequency selectivity. As we discussed earlier in the paper, implementing the scheme in [18] requires multiple radios in the relays.

8. CONCLUSION

We studied the problem of proportional fair scheduling in WiMAX relay networks. Unlike past work, we take into account the frequency selectivity, as well as multiuser diversity in our scheduling decisions. We show that the problem is NP-hard, and difficult to approximate. We provide a tight upper bound on the optimum and propose three heuristic algorithms, one for the generic OFDMA networks, and two for IEEE 802.16j standard settings. Through extensive simulations we demonstrate that using our proposed scheduling algorithms, relays can significantly improve the throughput and increase range effectively. Our algorithms are easy to implement and have a running time of less than 0.05 ms, thus suitable for WiMAX scheduling frame durations of $5 - 10$ ms.

Acknowledgement: We thank Kanthi Nagaraj of Bell Labs India for helping us with the fading module in our simulator.

9. REFERENCES

[1] *GLPK (GNU Linear Programming Kit), version 4.22*.

[2] M. Andrews. A survey of scheduling theory in wireless data networks. In *Proceedings of the 2005 IMA summer workshop on wireless communications*, 2005.

[3] M. Andrews and L. Zhang. Scheduling algorithms for multi-carrier wireless data systems. In *ACM Mobicom 2007*, September 2007.

[4] M. Charafeddine, O. Oymant, and S. Sandhu. System-level performance of cellular multihop relaying with multiuser scheduling. In *CISS*, March 2007.

[5] Y. W. Cheong, R. S. Cheng, K. B. Latief, and R. D. Murch. Multiuser ofdm with adaptive subcarrier, bit and power allocation. *IEEE Journal on Selected Areas in Communications*, October 1999.

[6] S. Deb, V. Mhatre, and V. Ramaiyan. WiMAX relay networks: Opportunistic scheduling to exploit multiuser diversity and frequency selectivity. *Bell Labs Technical Report*, Feb 2008.

[7] M. Ergen, S. Coleri, and P. Varaiya. QoS aware adaptive resource allocation techniques for fair scheduling in ofdma based broadband wireless access systems. *IEEE Tran. on Broadcasting*, Dec 2003.

[8] IEEE 802.16 task group. *Channel Models for Fixed Wireless Applications*, ieee 802.16.3c-01/29r4 edition, July 2001.

[9] IEEE 802.16e task group. *Air Interface for Fixed and Mobile Broadband Wireless Access Systems. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, 802.16e-2005 edition, February 2006.

[10] IEEE 802.16j task group. *Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Multihop Relay Specification*, 802.16j-06/026r4 edition, June 2007.

[11] O. Jo and D. Cho. Traffic adaptive uplink scheduling scheme for relay station in ieee 802.16 based multihop system. In *IEEE VTC,* 2004.

[12] R. Cohen L. Katzir. Computational analysis and efficient algorithms for micro and macro ofdma scheduling. In *IEEE Infocom 2008*.

[13] F. P. Kelly, A.K. Maulloo, and D.K.H. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, (49):237–252.

[14] D. Kivanc, G. Li, and H. Liu. Computationally efficient bandwidth allocation and power control for ofdma. *IEEE Transactions on Wireless Communications*, 2(6):1150–1158, November 2003.

[15] H. J. Kushner and P. A. Whiting. Convergence of proportional-fair sharing algorithms under general conditions. *IEEE Transactions on Wireless Communication*, 3(4):1250– 1259, July 2004.

[16] G. Narlikar, G. Wilfong, and L. Zhang. Designing multihop wireless backhaul networks with delay guarantees. In *IEEE Infocom 2006*.

[17] K. Navaie and Halim Yanikomeroglu. Multi-route and multi-user diversity in infrastructure-based multi-hop networks. *Cooperation in Wireless Networks: Principles and Applications, Editors: Frank H.P. Fitzek and Marcos D. Katz*, 2006.

[18] O. Oyman. OFDMA2A: A centralized resource allocation policy for cellular multi-hop networks. In *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov 2006.

[19] J. Padhye R. Draves and B. Zill. Routing in multi-radio, multi-hop wireless mesh network. In *ACM Mobicom*, September 2004.

[20] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall, 2001.

[21] W. Rhee and J. M. Cioffi. Increase in capacity of multiuser ofdm system using dynamic subchannel allocation. In *IEEE VTC*, 2000.

[22] Wimax forum. *Mobile WiMAX Part I: A Technical Overview and Performance Evaluation*, August 2006.

[23] Q. Wu and E. Esteves. The CDMA2000 high rate packet data system. *Chapter 4 of Advances in 3G Enhanced Technologies for Wireless Communications*. Editors: Jiangzhou Wang and Tung-Sang Ng.

[24] Y. Yu, S. Murphy, and L. Murphy. A clustering approach to planning base station and relay station locations in ieee 802.16j multi-hop relay networks. In *IEEE ICC*, 2008.