Low-overhead Scheduling Algorithms for OFDMA Relay Networks

(Invited Paper)

Karthikeyan Sundaresan, Xiaodong Wang, and Mohammad Madihian
Broadband and Mobile Networking, NEC Labs America

ABSTRACT

Enhanced data rates and connectivity are key requirements for providing ubiquitous mobile access in next-generation cellular networks. Relay-enabled cellular networks, marked by their adoption in IEEE 802.16j standard, have become a viable candidate in such an endeavor. Such relay networks not only provide multi-user and (OFDM) channel diversity gains that are available in conventional cellular systems, but also provide spatial reuse gains, arising from the simultaneous transmissions on different hops of the network on the same channel. However, the efficient exploitation of these gains, calls for intelligent scheduler design at the BS that must not only accommodate the multi-hop nature of the network, but also address the resulting significant overhead incurred in the form of feedback. In this work, we present *relay-assisted* scheduling algorithms that efficiently exploit the available diversity and spatial reuse gains at the cost of *minimal feedback overhead*. The proposed solutions improve performance over conventional approaches by over 50% along with a scalable feedback overhead that grows only with the number of relays in the network and not with the number of users.

Categories and Subject Descriptors

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

General Terms

Algorithms, relays, diversity, spatial reuse, scheduling, performance

1. INTRODUCTION

There has been a significant rise in the popularity of real-time multi-media services such as streaming audio and video, video-on-demand, IPTV, etc. This coupled with the advancement of various wireless access technologies, has ushered in an era of ubiquitous access to various forms of data and media content. However, the current cellular systems are not sufficiently equipped to meet the requirements of either ubiquitous coverage or bandwidth-intensive real-time applications. Given the large capital investment in cellular network infrastructure, an efficient way to meet the future demands would be to reuse the existing infrastructure but upgrade it with appropriate functionalities. One such popular approach is *relay-enabled* cellular networks, whereby less sophisticated relay stations (RS) are introduced into the cellular network to help in the transport of data between the base station (BS) and the mobile stations (MS). Such relay-enabled networks have been shown to provide improved capacity and coverage over the conventional cellular networks [1, 2], contributing to their adoption in the IEEE 802.16j relay task group, with OFDM as the air-interface technology. Since most of their envisioned applications (access inside transportation vehicle, buildings, etc.) follow the two-hop network model (mandatory in 802.16j), this forms our focus in this work.

These networks are different from the conventional multi-hop and cellular networks and hence require unique optimizations. In OFDM cellular networks, the single-hop nature allows for efficient, centralized exploitation of diversity (multi-user and multi-channel) gains at the BS [3, 4]. However, they do not provide enhanced connectivity or spatial reuse (on the same channel) to provided improved data rates. On the other hand, multi-hop networks allow for spatial reuse on the different hops [5], but the significant coordination and overhead arising from the multi-hop nature prevents efficient exploitation of diversity gains in a centralized fashion. The relay networks have the potential to provide the *best of both worlds*. The two-hop nature allows for centralized exploitation of diversity gains at the BS. Further, it also allows for spatial reuse between the two hops. However, in order to effectively leverage the diversity and spatial reuse gains, we need efficient scheduling algorithms. Further, the feedback overhead arising from the two hops must be contained so that it does not outweigh the benefits of increased network capacity resulting from spatial reuse.

We consider two popular models for RS. (i) Interference-unaware RS: The RS and the network do not support interference estimation and reporting functionalities. This provides no room for spatial reuse, although diversity gains resulting from the two-hops are significantly higher than in conventional cellular networks. Every time slot is divided into two sub-slots, one for each hop transmission in this model. This model has been considered in several proposals in 802.16j [6]. (ii) Interference-aware RS: The RS and the network support interference estimation and reporting functionalities. This is the most sophisticated model, wherein both diversity and spatial reuse gains from the two hops can be leveraged. Our contributions can be summarized as follows.

- A low-feedback, near-optimal scheduling algorithm for exploiting diversity (multi user/channel) in the interference-unaware model. A $\frac{1}{2}$-approximation algorithm is also proposed for the network bottleneck case under QoS constraints;

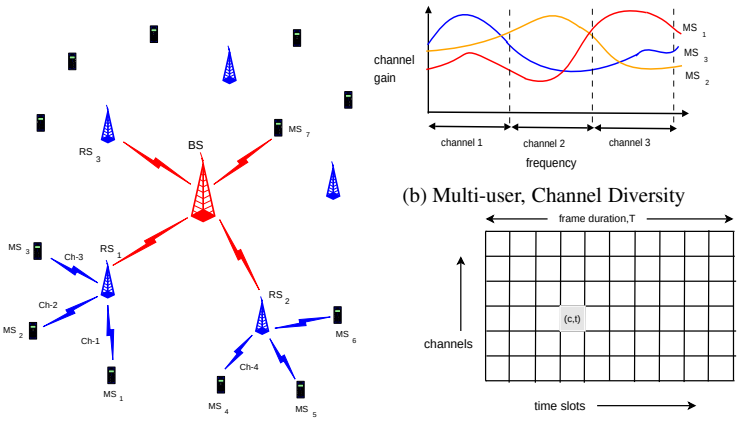
WICON'08, November 17-19, 2008, Maui, Hawaii, USA.

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

- A low-feedback, efficient scheduling algorithm for exploiting diversity and spatial reuse in the interference-aware model; and
- All the proposed algorithms incur a scalable feedback overhead that grows only with the number of relay stations and not with the number of users in the network.

The fact that RS have more information on the network state than the BS is exploited by the algorithms to allow the RS assist the BS in its scheduling functionality and reduce the feedback overhead in the process without compromising in performance.

The remainder of the paper is organized as follows. Related work along with network and scheduling models are presented in Section 2. Low feedback scheduling algorithms to exploit diversity within and across hops are presented in Section 3. Section 4 presents the low feedback scheduling algorithm to exploit spatial reuse in addition to diversity gains. The proposed solutions are evaluated in Section 5, followed by concluding remarks in Section 6.



(a) Network Model

(b) Multi-user, Channel Diversity

(c) Frame Structure

Figure 1: System Model and Gains

2. SYSTEM DESCRIPTION

2.1 Related Work

Several works [1, 2] have investigated the potential of relay-enabled wireless networks (cellular and WLANs) to provide improved coverage and capacity. These networks have also gained attention in the standards community (IEEE 802.16j) as well as from the industry.

Scheduling [7, 8] has been identified to be an important aspect critical to leveraging the potential benefits of these networks. However, most of these works focus on link level performance and do not exploit spatial reuse that is available at a network level. Further, they do not consider a multiple channel OFDM network (channel diversity) and QoS constraints, which complicate scheduling decisions and increases overhead with the possibility of multiple users operating in parallel. The works on OFDM scheduling in conventional cellular systems [4, 3] cannot be directly applied to two-hop cellular networks, where the network structure is different and spatial reuse forms an important component. There have been some works [9, 10] 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. Once again the focus is on a link-level performance, failing to exploit spatial reuse. Our recent work [11] looks at the problem of optimal diversity scheduling in two-hop relay networks. However, none of these works consider the increase in feedback overhead resulting from multiple channels and hops, and interference arising from spatial reuse. Thus, our focus in this work is to design efficient scheduling algorithms that exploit diversity benefits (across users, relays and hops) and spatial reuse gains (available across the hops), while incurring *reduced feedback overhead*.

2.2 Network Model

We consider a downlink OFDMA-based, relay-enabled, two-hop wireless network as shown in Figure 1(a). 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 [12]. 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 1(b). Multipath fading and user mobility result in independent fading across users for a given channel, 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).

The spatial seperation 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.

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 1(c). 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

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

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 proposed solutions.

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 [13]. Throughput and fairness are obtained by performing scheduling such that it maximizes aggregate network utility: $\max\left\{\sum_{k=1}^{K}\beta_k U_k\right\}$, where U_k represents the utility of user k for a certain achieved (two-hop) throughput and β_k represents the priority weight of its QoS class. For concave, continuously differentiable and increasing utility functions, the system can be shown to converge to the optimum if the scheduler's decisions at each time slot are based on the maximum marginal utility. Thus, the schedule (S_{max}) for each time slot is given by: $S_{\max} = \arg\max_S\left\{\sum_{k\in S}\Delta U_k\right\}$, where ΔU_k denotes the marginal flow (two-hop) utility of user k in a feasible schedule S in the presence of relays. The nature of utility function determines the fairness model achieved in the system. For proportional fairness, $U(r) = \log r$ and the corresponding marginal utility (ΔU_k) of a user depends on both its average throughput $\bar{r}_k$ as well as its instantaneous rate ($\Delta U_k = \frac{r}{\bar{r}_k}$). However, the instantaneous rate now corresponds to the *two-hop flow* rate, which in turn is determined by the instantaneous *effective* rate on the relay and access hops combined.

If $r_{k,n}^{rel}$ and $r_{k,m}^{acc}$ are the bit-rates obtained for a user (flow) k on the relay and access links on channel n and m respectively, then the effective end-to-end rate of the two-hop transmission is,

$$\frac{1}{r_{k,n,m}^{eff}} = \frac{1}{r_{k,n}^{rel}} + \frac{1}{r_{k,m}^{acc}} \Rightarrow r_{k,n,m}^{eff} = \frac{r_{k,n}^{rel}\cdot r_{k,m}^{acc}}{r_{k,n}^{rel} + r_{k,m}^{acc}}$$

Thus, it captures the transmission delay incurred by the packet in being transported from BS to the MS through the two hops and hence provides a measure of the end-to-end rate for the packet transport. This is referred to as the *effective multi-hop throughput*. The motivation for considering this metric is its adoption as the throughput evaluation metric in IEEE 802.16j standard that is being ratified.

While there exist several utility-based schedulers for the one-hop flows, the challenge arises in exploiting the diversity and spatial reuse gains available with two-hop flows. Further, while spatial reuse is possible across hops and within access links, it is not possible within relay links (due to the common transmitter/receiver (BS) for relay hop transmissions. As a result, the relay hop forms the network capacity bottleneck. Consequently, while channel state feedback from MS to RS on the access links is feasible, further propagation of such information to BS constitutes significant feedback overhead on the relay links and must hence be kept low. For eg. if there are R relays, K users ($\frac{K}{R}$ users per relay) and N channels, then the amount of feedback required on the relay hop is $O(KN)$. Thus, the feedback grows with the number of users in the system and is hence not scalable. Designing solutions with low feedback overhead for relays form an important aspect of the upcoming 802.16m standard. Hence, the main focus of this work is to retain a performance close to that of the full feedback scheme albeit at the cost of a feedback that scales only with the number of relays in the network, namely $O(RN)$. Note that R does not grow with K in the system. This keeps the feedback overhead contained on the relay hop, thereby allowing the relay hop resources be efficiently used towards improving the system throughput.

3. INTERFERENCE-UNAWARE RS

The RS are assumed to be capable of modifying the data frame that is passed on from the BS and destined to an associated MS. This gives the RS the potential to switch the channel on which data was received for a particular user on the relay link to a different channel on the access link, thereby allowing for diversity both within and across hops. However, no interference estimation/reporting is considered and hence no exploitation of spatial reuse is possible at the BS. Under the given RS model along with the half-duplex constraint of RS, a joint two-slot flow schedule of the relay links followed by the access links can be obtained that exploits diversity both within and across hops. The marginal utility of a user (flow) k on a relay-access channel pair (n,m) acrosss two hops (slots) is given by $\frac{\beta_k r_{k,n,m}^{eff}}{\bar{r}_k}$. The scheduling problem is now equivalent to: Find a subset of users and a corresponding assignment of N sub-channels on relay links to N sub-channels on access links such that aggregate marginal utility is maximized with no channel on either hops being assigned to more than one user. The relay and access sub-channels are assigned to flows (users) in pairs with a user capable of being assigned multiple channel pairs. More formally, we have

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

$$\sum_{k=1}^{K}\sum_{n=1}^{N} I_{k,n,m}(t) \le 1, \quad \forall m \qquad \sum_{k=1}^{K}\sum_{m=1}^{N} I_{k,n,m}(t) \le 1, \quad \forall n \quad (1)$$

where $I_{k,n,m}(t) \in \{0,1\}$, is a binary function capturing the assignment of (relay,access) channel pair (n,m) to user k in slot t. The above problem is solved optimally by solving an equivalent maximum utility bipartite matching problem as follows.

- 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 links with $|V_1| = |V_2| = N$. The edge set E corresponds to $|N|^2$ edges connecting all possible pairs of vertices in the two sets.
- The weight on each of the edges W_{ij} depends on the user to whom the channel pair is assigned. Every weight carries two attributes, (w_{ij}, u_{ij}), where u_{ij} and w_{ij} correspond to the user assigned and its marginal utility respectively. The attribute w_{ij} of an edge is now obtained by the maximum of the marginal utilities among all possible assignments of users to the channel pair under consideration. Using marginal utilities as the weights takes into account the average throughput of users and hence fairness.
  - $w_{ij} = \max_k\{\beta_k \Delta U_k\} = \max_k\left\{\frac{\beta_k r_{k,i}^{rel} r_{k,j}^{acc}}{\bar{r}_k\left(r_{k,i}^{rel}+r_{k,j}^{acc}\right)}\right\}$
  - $u_{ij} = \arg\max_k\left\{\frac{\beta_k r_{k,i}^{rel} r_{k,j}^{acc}}{\bar{r}_k\left(r_{k,i}^{rel}+r_{k,j}^{acc}\right)}\right\}$
- It is now easy to see that finding the maximum weight bipartite matching on G now provides the set of N channel pair assignments on the relay and access links that bring in the maximum marginal utility. Further, the second attribute of the edges present in the maximum matching provide the set of MS and associated RS 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 and we use the Hungarian algorithm [14] of finding augmenting paths for our design.

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

3.1 Addressing Scalability of Overhead

While the above algorithm provides the optimal schedule at every slot, feedback on all the sub-channels is required from both the access and relay links for every user. While feedback on the access channels (from MS) can be obtained at the RS, this has to be propagated to the BS in addition to the relay channel feedback from RS. This requires significant feedback overhead on the relay hop, which is already a bottleneck and must hence be addressed. If B_r corresponds to the number of bits used to feedback rate information on a given sub-channel, then the feedback overhead incurred on the relay links is $(K+R)\cdot N\cdot B_r$, which is $O((K+R)\cdot N)$ and hence grows with the number of users, bringing down the relay hop (network) capacity, which is not desirable (scalable). This is because, for a given channel on the access link, the optimal user at the RS could be different depending on the specific channel chosen in the relay link. The choice of relay channel would in turn depend on the global decision at the BS, which is not known apriori at the RS. To address this issue, we now present some properties and subsequently exploit them to propose a scalable feedback scheduler that retains close-to optimal performance. For simplicity, hereafter we incorporate β_k into the average throughput of the user, $\bar{r}_k \leftarrow \frac{\bar{r}_k}{\beta_k}$.

LEMMA 1. *Consider two users i and j associated with the same RS and let m be an access channel. Let the one-hop marginal utilities of the users on the access links be such that,*

$$\frac{r_{i,m}^{acc}}{\bar{r}_i} \geq \frac{r_{j,m}^{acc}}{\bar{r}_j}$$

with user i belonging to the feedback set. Now, user j's feedback can be eliminated, irrespective of the relay channel chosen, if any of the following conditions are true:

(1) $r_{i,m}^{acc} < r_{j,m}^{acc}$, *(2)* $\bar{r}_i < \bar{r}_j$, *(3)* $r_{CO}^{rel} = \frac{r_{i,m}^{acc}\left(1-\frac{\bar{r}_j}{\bar{r}_i}\right)}{\frac{r_{i,m}^{acc}}{r_{j,m}^{acc}}\cdot\frac{\bar{r}_j}{\bar{r}_i}-1} < r_{min}^{rel}$

PROOF. Since both users are associated with the same relay, $r_{i,n}^{rel} = r_{j,n}^{rel} = r_n^{rel}$, for any relay channel n. For user j to provide a higher effective marginal utility, we need

$$\frac{r_{i,m}^{acc}\cdot r_n^{rel}}{\bar{r}_i\left(r_{i,m}^{acc}+r_n^{rel}\right)} \leq \frac{r_{j,m}^{acc}\cdot r_n^{rel}}{\bar{r}_j\left(r_{j,m}^{acc}+r_n^{rel}\right)}$$

Rearranging, we have,

$$r_n^{rel} \leq \frac{r_{i,m}^{acc}\left(1-\frac{\bar{r}_j}{\bar{r}_i}\right)}{\frac{r_{i,m}^{acc}}{r_{j,m}^{acc}}\cdot\frac{\bar{r}_j}{\bar{r}_i}-1} = r_{CO}^{rel} \qquad (2)$$

This provides the set of relay channel rates (below the cross-over rate, r_{CO}^{rel}) for which user j will provide a higher effective marginal utility than i. Given $\frac{r_{i,m}^{acc}}{r_{j,m}^{acc}}\cdot\frac{\bar{r}_j}{\bar{r}_i} \geq 1$, for the set of relay rates to be feasible, we need $\frac{\bar{r}_j}{\bar{r}_i} \leq 1$ and $\frac{r_{i,m}^{acc}}{r_{j,m}^{acc}} \geq 1$. Also, we need the cross-over rate to be larger than the minimum rate available on the relay channels, $r_{CO}^{rel} \geq r_{min}^{rel}$. □

As a corollary, if conditions 1 and 2 are false and $r_{CO}^{rel} \geq r_{max}^{rel}$, then user i can be removed from the feedback list and replaced by user j. As an extension to lemma 1, we also have,

LEMMA 2. *Given three users, i, j, and k with feedback from users i and k on access channel m and $\frac{r_{i,m}^{acc}}{\bar{r}_i} \geq \frac{r_{j,m}^{acc}}{\bar{r}_j} \geq \frac{r_{k,m}^{acc}}{\bar{r}_k}$, then user j's feedback on access channel m can be eliminated if,*

$$r_{j,m}^{acc} \leq \frac{\left(\frac{1}{\bar{r}_k}-\frac{1}{\bar{r}_i}\right)\bar{r}_j\cdot r_{i,m}r_{k,m}}{(r_{i,m}-r_{k,m})-\bar{r}_j\left(\frac{r_{i,m}}{\bar{r}_i}-\frac{r_{k,m}}{\bar{r}_k}\right)} = r_{th}^{acc} \qquad (3)$$

PROOF. Relay channel rates for which user j will provide a higher marginal utility than user i is given by,

$$r_n^{rel} \leq \frac{r_{i,m}^{acc}\left(1-\frac{\bar{r}_j}{\bar{r}_i}\right)}{\frac{r_{i,m}^{acc}}{r_{j,m}^{acc}}\cdot\frac{\bar{r}_j}{\bar{r}_i}-1} = \frac{r_{i,m}^{acc}\cdot r_{j,m}^{acc}\cdot(\bar{r}_i-\bar{r}_j)}{r_{i,m}^{acc}\cdot\bar{r}_j - r_{j,m}^{acc}\cdot\bar{r}_i}$$

Similarly, relay channel rates for which user j will provide a higher marginal utility than user k is governed by,

$$r_n^{rel} \geq \frac{r_{k,m}^{acc}\cdot r_{j,m}^{acc}\cdot(\bar{r}_j-\bar{r}_k)}{r_{j,m}^{acc}\cdot\bar{r}_k - r_{k,m}^{acc}\cdot\bar{r}_j}$$

For user j to be eliminated, the above two inequalities must provide an infeasible relay rate region. This results in ,

$$\frac{r_{i,m}^{acc}\cdot r_{j,m}^{acc}\cdot(\bar{r}_i-\bar{r}_j)}{r_{i,m}^{acc}\cdot\bar{r}_j - r_{j,m}^{acc}\cdot\bar{r}_i} \leq \frac{r_{k,m}^{acc}\cdot r_{j,m}^{acc}\cdot(\bar{r}_j-\bar{r}_k)}{r_{j,m}^{acc}\cdot\bar{r}_k - r_{k,m}^{acc}\cdot\bar{r}_j}$$

Simplifying, we obtain the desired result. □

Remarks: Given a set of users associated with a RS and arranged in the decreasing order of access hop marginal utilities, applying lemmas 1 and 2 results in a significantly reduced feedback list that retains optimal performance. To see this, if the difference in the average throughputs of users is small to moderate, r_{CO}^{rel} in equation 2 tends to a small value potentially lesser than r_{min}^{rel}, thereby requiring feedback only from the higher (access link) marginal utility users. In the limiting case, with $\frac{\bar{r}_j}{\bar{r}_i} \rightarrow 1$, then $r_{CO}^{rel} \rightarrow 0$ requiring only the highest marginal utility user's feedback. Similarly, when the difference in the average throughputs is moderate to large, r_{CO}^{rel} tends to a large value, potentially larger than r_{max}^{rel}, thereby requiring feedback only from the smaller marginal utility users. This would correspond to the smallest marginal utility user in the limiting case when $\frac{\bar{r}_i}{\bar{r}_j} \rightarrow \frac{r_{i,m}^{acc}}{r_{j,m}^{acc}}$, where $r_{CO}^{rel} \rightarrow \infty$. Applying lemmas 1 and 2 automatically takes into account the distribution of user throughputs in determining the minimal set of users (for feedback) for optimal performance, which in turn is a single element for the limiting cases and a small set of elements otherwise. Hence, to incorporate small to large variations in user throughputs, providing the two extreme (largest and smallest marginal utility) users from this final reduced feedback list is sufficient to provide near-optimal performance, while keeping the feedback overhead scalable. This is verified in our evaluations as well.

Using the above lemmas and assistance from relays, the reduced feedback scheduling algorithm at BS for exploiting diversity both within and across hops (MAXDIV-RF) is presented in Algorithm 1. From steps 5 and 6, it can be seen that for every relay, two rate elements and two index elements are required as feedback on each of the access channels, while a single rate element is required on each of the relay channels. Thus, if B_a represents the number of bits used to indicate user id ($B_a < B_r$), then the *feedback overhead on the relay links incurred by MAXDIV-RF is now given by, $R\cdot N\cdot(3B_r+2B_a) = O(R\cdot N)$ that scales only with the number of relays, which is a small, fixed parameter unlike the number of users*. The benefits of MAXDIV-RF over a feedback scheme of choosing two users providing the highest (access hop) marginal utilities, is presented in the evaluations section.

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

Algorithm 1 Max. Diversity at Reduced Feedback: MAXDIV-RF

1: Each RS q assigns all its K_q users to potential feedback set L_m $\forall$ access channel m.
2: $\forall m$, each RS q arranges users from the feedback set in decreasing order of access link marginal utility: $\frac{r^{acc}_{1,m}}{\bar{r}_1} \geq \frac{r^{acc}_{2,m}}{\bar{r}_2} \geq \cdots \geq \frac{r^{acc}_{K_q,m}}{\bar{r}_{K_q}}$. A mapping of the sort list index to actual user index is maintained.
3: $\forall \ell \in [1, K_q]$, RS q removes the users (v) following ℓ that satisfy atleast one of the conditions in lemma 1, reducing the set L_m to $K_{q,1}$ users.
4: RS q starts with the last (lowest) utility user and moves iteratively towards the first user. $\forall j$, RS determines if its feedback is required with respect to its preceding and succeeding users in the feedback list using lemma 2. If not, user j is removed from potential feedback, reducing the feedback set L_m further to $K_{q,2}$ users.
5: RS q sends access rate information, user id for its two extreme users from final set L_m. Let $u^{max}_{q,m} = \max_{k \in L_m} \left\{ \frac{r^{acc}_{k,m}}{\bar{r}_k} \right\}$, $u^{min}_{q,m} = \min_{k \in L_m} \left\{ \frac{r^{acc}_{k,m}}{\bar{r}_k} \right\}$.
$F^{acc}_{q,m} = \left\langle u^{max}_{q,m}, u^{min}_{q,m}, \arg\left\{u^{max}_{q,m}\right\}, \arg\left\{u^{min}_{q,m}\right\} \right\rangle, \forall m.$
6: RS q sends relay rate information: $F^{rel}_{q,n} = \frac{r^{rel}_{q,n}}{\bar{r}_k}, \forall n.$
7: BS constructs bipartite graph G with:
$w_{ij} = \max_q \left\{ \frac{F^{rel}_{q,i} F^{acc}_{q,j}(1)}{\left(F^{rel}_{q,i} + F^{acc}_{q,j}(1)\right)}, \frac{F^{rel}_{q,i} \cdot F^{acc}_{q,j}(2)}{\left(F^{rel}_{q,i} + F^{acc}_{q,j}(2)\right)} \right\}$, $u_{ij} = \arg\{w_{ij}\}$.
8: BS runs Hungarian(G) to obtain the final schedule.

3.2 Scheduling with QoS constraint

In exploiting diversity across hops, optimal solution has been possible since users were allowed to be assigned multiple channels. In the presence of QoS constraints, if a limit is placed on the number of channels (C) that can atmost be assigned to a user, then the problem immediately becomes hard to solve.

The scheduling problem is now equivalent to SQC: Find a subset of users and a corresponding assignment of N sub-channels on relay links to N sub-channels on access links such that aggregate marginal utility is maximized with no channel on either hops being assigned to more than one user and no user being assigned more than C channels on any hop. The formulation in 1 remains the same with the addition of the following constraint.

$$\sum_{m=1}^{N} \sum_{n=1}^{N} I_{k,n,m}(t) \leq C, \quad \forall k$$

THEOREM 1. *Problem SQC is NP-hard.*

PROOF. This can be established by a polynomial-time reduction from maximum weight *3-dimensional matching*, which is known to be NP-hard [15]. The *3-dimensional matching* problem is stated as follows.

Given a set $T \subseteq V_1 \times V_2 \times V_3$, where V_1, V_2, and V_3 are disjoint, and a weight function, $w_{i,j,k} \geq 0, \forall\, i \in V_1,\, j \in V_2$, and $k \in V_3$, find a maximum weight matching (M) for T, i.e. $M \subseteq T$ such that no elements in M agree in any coordinate and the aggregate weight of the matching is maximum.

Construct the following tripartite graph $G = (V_1 \times V_2 \times V_3, E)$. V_1, V_2 and V_3 represent the set of N relay channels, N access channels and K users, each user being replicated C times, resulting in KC total users. The weight of an edge $(i, j, k) \in E, i \in V_1, j \in V_2, k \in V_3$ is computed using the marginal utilities as follows.

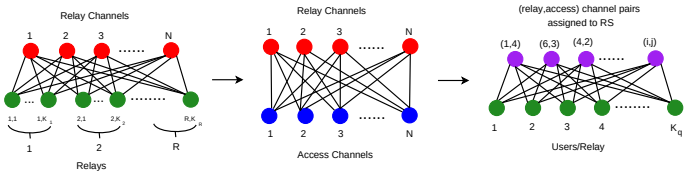


Figure 2: 2D reduction under bottleneck condition

$$w_{i,j,k} = \frac{1}{\bar{r}_k} \cdot \frac{r^{rel}_{k,i} \cdot r^{acc}_{k,j}}{r^{rel}_{k,i} + r^{acc}_{k,j}}$$

Replicated users will have the same weight for a given channel pair as the original user. Now, finding the maximum weight 3-dimensional matching on G, yields the schedule (assignment) that solves X, with no more than C channel pairs being assigned to any user. □

The problem can however be solved efficiently for decomposable weight functions, $w_{ijk} = w_i w_j w_k$, where polynomial-time solutions exist and for the maximum cardinality version of the problem, where some good approximations exist. Since we are interested in the much harder maximum weight version and our non-decomposable weight function does not fall into any of the well-known categories, this makes it all the more difficult to construct efficient approximations. However, we provide a $\frac{1}{2}$-approximation algorithm under the restricted case of bottleneck conditions, where either the relay or the access hop forms the network bottleneck.

The relay hop forming a bottleneck is in fact very common and practical given the heterogeneity in technologies possible across the two hops; a popular example being the low bandwidth WAN links on the relay hop and the high bandwidth WLAN links on the access hop, with RS serving as access points. We now present an algorithm that exploits this bottleneck nature of the network to decompose the 3D matching into a series of 2D matchings and hence solve it with worst case guarantees. The relay-assisted algorithm satisfying the QoS constraint (QOS-RF) is presented in Algorithm 2 for $C = 1$ and illustrated in Figure 2.

Algorithm 2 Diversity + QoS at Reduced Feedback: QOS-RF

1: BS obtains a maximum weight utility matching (MWUM) B_1 on $G = (V_1 \times V_2, E)$, matching the set of relay channels (V_1) to the set of relays (V_2), with each relay replicated K_q times (# users associated with RS q) and $w_{i,k} = \frac{\max_j \left\{ \min\left\{ r^{rel}_{k,i}, r^{acc}_{k,j} \right\} \right\}}{\bar{r}_k}$.
2: BS obtains a MWUM B_2, matching the set of relay channels to access channels, while satisfying the assignment of relay channels to RS from B_1, with $w_{i,j} = \max_k \{w_{i,j,k}\}$, where $w_{i,j,k} = \frac{1}{\bar{r}_k} \cdot \frac{r^{rel}_{k,i} \cdot r^{acc}_{k,j}}{r^{rel}_{k,i} + r^{acc}_{k,j}}$, if $(i, relay(k)) \in B_1$ and 0 otherwise.
3: Each RS q obtains a MWUM B_3, matching the set of set of (relay,access) pairs assigned to q from B_2 to its associated users, with $w_{ij,k} = \frac{1}{\bar{r}_k} \cdot \frac{r^{rel}_{k,i} \cdot r^{acc}_{k,j}}{r^{rel}_{k,i} + r^{acc}_{k,j}}$, if $(i, j) \in B_2$ $\&(i, relay(k)) \in B_1$, and 0 otherwise.
4: B_3 provides the required schedule satisfying the QoS constraint.

THEOREM 2. *QOS-RF is a $\frac{1}{2}$-approximation algorithm for SQC under the bottleneck condition.*

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

PROOF. Let i, j, k belong to the set of relay channels, access channels and users respectively. Define, $w_{i,k} = \frac{r^{rel}_{i,k}}{\bar{r}_k}$, $w_{j,k} = \frac{r^{acc}_{j,k}}{\bar{r}_k}$ and $w_{i,j,k} = \frac{1}{\frac{1}{w_{i,k}}+\frac{1}{w_{j,k}}} = \frac{1}{\bar{r}_k} \cdot \frac{r^{rel}_{k,i} \cdot r^{acc}_{k,j}}{r^{rel}_{k,i}+r^{acc}_{k,j}}$. Assume $w_{i,k} \leq w_{j,k}$ without loss of generality. Now, $\min\{w_{i,k}, w_{j,k}\} = w_{i,k}$. Hence,

$$w_{i,j,k} = \frac{1}{\frac{1}{w_{i,k}} + \frac{1}{w_{j,k}}} \geq \frac{1}{\frac{2}{w_{i,k}}} = \frac{w_{i,k}}{2}$$

On the other hand,

$$w_{i,j,k} = \frac{1}{\frac{1}{w_{i,k}} + \frac{1}{w_{j,k}}} \leq \frac{1}{\frac{1}{w_{i,k}}} = w_{i,k}$$

Thus, we have $\frac{\min\{w_{i,k}, w_{j,k}\}}{2} \leq w_{i,j,k} \leq \min\{w_{i,k}, w_{j,k}\}$. This in turn implies that if a problem with objective function $w_{i,j,k}$ is replaced by $\min\{w_{i,k}, w_{j,k}\}$, then an algorithm solving the latter optimally will provide a $\frac{1}{2}$-approximation solution to the former in the worst case, resulting in an aggregate utility that is atleast half that of the optimal.

While even the modified objective version is difficult to solve in the general case, however under the bottleneck condition (say, relay hop forming bottleneck), we now have $\min\{r^{rel}_{k,i}, r^{acc}_{k,j}\} = r^{rel}_{k,i}$, $\forall j$. This allows $w_{i,k}$ in step 1 of algorithm QOS-RF to reduce to $\frac{r^{rel}_{k,i}}{\bar{r}_k}$, thereby removing the dependency on j, with steps 1 and 3 ensuring the QoS constraint. Thus, QOS-RF now solves the modified objective function optimally and hence provides a $\frac{1}{2}$-approximation to the original objective function. This also applies to the case where the access hop forms the bottleneck. □

Note that, subject to the assignment from step 1, any feasible assignment of access channels to users satisfying the QOS constraint would retain the same worst case guarantee. However, to provide a much better performance in the average case, 2D matchings in steps 2 and 3 are adapted to maximize the aggregate marginal utility subject to the assignment from step 1. Further, under the *relay bottleneck* condition, the feedback information needed by the BS for steps 1 and 2 are the same as that in MAXDIV-RF. Hence, *QOS-RF also incurs an overhead of $O(R \cdot N)$ that scales only with the number of relays*. Also, QOS-RF can be extended to $C > 1$ case with the same worst case guarantee by allowing the number of users in steps 1 and 3 to be replicated C times. The algorithm also provides good average case performance in general network conditions as demonstrated in evaluation results, which can be attributed to steps 2 and 3 of the algorithm.

4. INTERFERENCE-AWARE RS

Building on the diversity scheduler, we now design a scheduler for leveraging both diversity and spatial reuse to improve network capacity. In the previous model, we assumed that the RS did not support interference reporting and hence exploited only diversity gains, with the relay and the access hops being scheduled sequentially without exploiting spatial reuse. We now consider interference-aware RS and hence leverage the potential spatial reuse *across* relay and access hops to help improve network capacity. Note that, while the spatial separation between RS allows for spatial reuse even *within* access links, there is not much room for its exploitation especially when the relay hop forms the bottleneck (as demonstrated in evaluations). Thus, we restrict our initial focus to exploiting spatial reuse only *across* relay and access links and outline how it can be extended to leverage spatial reuse within access links later. 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

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

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

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

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 [5]. (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-RF

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 our proposed low feedback diversity scheduler, MAXDIV-RF 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 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 partition the set of RS based only on the traffic load in the network.

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

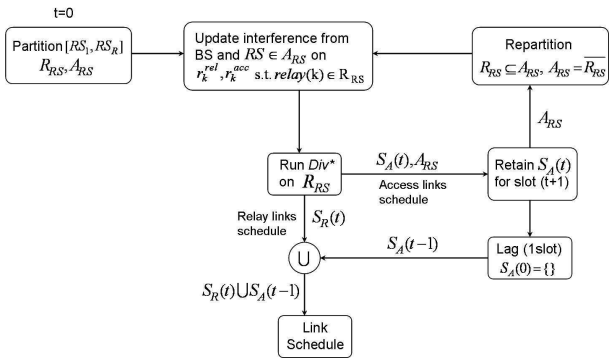


Figure 3: Algorithm

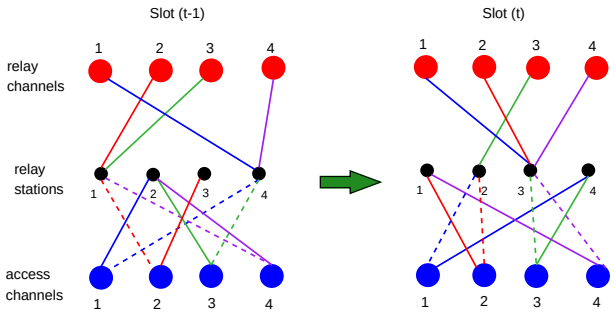


Figure 4: Illustration for exploiting spatial reuse and diversity

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. 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,

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

The following dynamic program yields the desired partition.

$$\begin{aligned} (q,d)^* &= \arg\min L[q,d] \\ L[q,d] &= \max\{\ell_{q,d}, W - \ell_{q,d}\}, \quad \forall (q,d) \\ \ell_{q,d} &= \ell_{q-1,d+1} - \ell_{q-1,1} \end{aligned}$$

The base cases from which the cost of the larger partitions can be 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 $O(R^2K)$ 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 at 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 at 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 $\overline{R_{RS}}$ ($\{RS_1, RS_4\}$). Our low feedback diversity scheduler (MAXDIV-RF) 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 MAXDIV-RF, 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 ($\chi_{BS \to k,n}$) in their instantaneous access rate feedback without (a priori) knowledge of the specific relay link to be scheduled on the same channel: $r_{k,n}^{acc} = \log(1 + \frac{P_{k,n}}{N_{k,n} + \chi_{BS \to k,n}})$, $\forall n$, 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 $S_A(t-1)$ 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 from $RS_j \in A_{RS}$ at $RS_q \in R_{RS}$ ($\chi_{j \to q,n}$) is then incorporated in the relay channel feedback as, $r_{q,n}^{rel} = \log(1 + \frac{P_{q,n}}{N_{q,n} + \sum_{j \in Access\_RS} \chi_{j \to q,n} B_{j,n}})$, $\forall n$, where $B_{j,n} = 1$ if $BM_{acc}(n) = RS_j$, and 0 otherwise.

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938

An interference-aware MAXDIV-RF 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.

4.2 Feedback Overhead

A typical interference reporting scheme as considered in current standards would require all the MS to send interference information from BS on all access channels, and the RS to send interference information from all other RS on all the relay channels. This interference information alone would incur an overhead of $(K + R \cdot (R-1)) \cdot N \cdot B_r$. In addition, the feedback required by conventional matching would incur $(K+R) \cdot N \cdot B_r$, resulting in a total feedback overhead of $(2K + R^2) \cdot N \cdot B_r$, which is $O\left((K+R^2) \cdot N\right)$. However, in SR+DIV-RF, the MS and RS incorporate interference directly in their rate estimates and need to feedback no additional interference information. However, to aid the RS on relay links in the estimation of interference from other RS on access links, a N-field bitmap is sent by the BS, requiring an overhead of $N \log_2 R$. This coupled with the feedback required by *MAXDIV-RF, results in a net feedback of* $R \cdot N \cdot (3B_r + 2B_a) + N \log_2 R$*, which is* $O(R \cdot N)$. Thus, the feedback does not scale with the number of users while both spatial reuse and diversity gains are exploited.

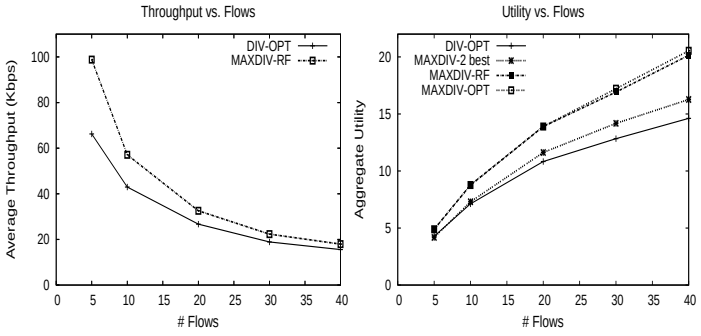
In summary, the benefits of proposed algorithms (MAXDIV-RF, QOS-RF, SR+DIV-RF) are:

	Performance		Feedback Overhead	
	Solvability	Proposed	Optimal	Proposed
Diversity	P	Near-optimal	$O((K+R)N)$	$O(RN)$
Diversity + QoS	NP-hard	$\frac{1}{2}$-approx	$O((K+R)N)$	$O(RN)$
Reuse + Diversity	NP-hard	Small gap	$O((K+R^2)N)$	$O(RN)$

5. PERFORMANCE EVALUATION

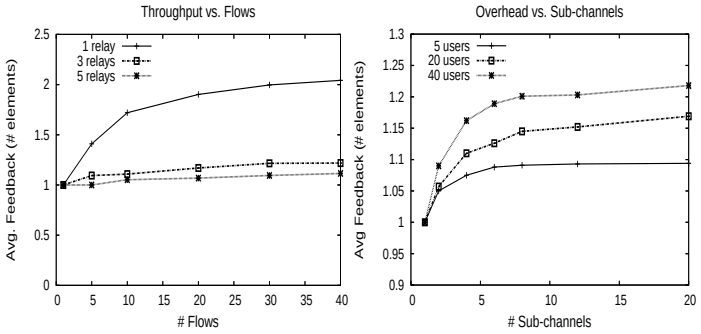
An event-driven packet level simulator written in C++, named *queuing network simulator* [16] is considered for evaluation of the proposed solutions. A single cell relay-enabled OFDM downlink system is considered. The extended radius of the cell is assumed to be about 600m. RS are distributed uniformly within a region of $250m \leq r \leq 350m$. The wireless links incorporate path loss, log-normal shadowing 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 consider 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 and sub-channels vary from [1,40], [1,10] and [1,20] respectively. 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. Results are measured either as a function of increasing users or sub-channels (bandwidth). Since the significant reduction in feedback overhead of the proposed schedulers has already been established, we now evaluate only their average per-user end-end throughput and aggregate network utility under the two RS models.



a) Throughput vs. Users: (x,8C,5R) b) Utility vs. Users: (x,8C,5R)

Figure 5: Exploiting diversity within and across hops



a) 20 sub-channels b) 3 relays

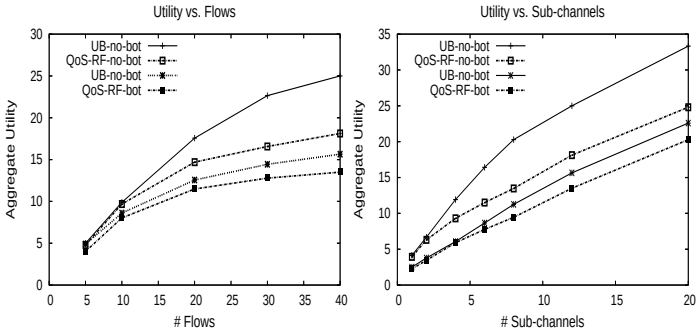
Figure 6: Feedback Overhead

5.1 Interference-unaware RS

The performance of MAXDIV-RF is evaluated against an optimal scheme DIV-OPT that exploits diversity only within hops (and not across hops) by using the same channel for transmission on both hops for a given user. Figure 5(a) presents throughput results as a function of increasing users with sub-channels and relays fixed at eight and five respectively. It can be seen that channel switching across hops helps leverage additional diversity gains to provide significant performance improvements of about 50% when channel diversity dominates over multi-user diversity (number of users being small/comparable to channels). To account for both throughput and fairness, the aggregate utility results are presented in Figure 5(b) as a function of increasing users. Utility gains of 30-50% are obtained and are more at increased number of users due to the concave nature of the utility function resulting in higher gains at lesser user throughput (higher load). In addition to DIV-OPT and MAXDIV-RF, optimal MAXDIV (full-feedback) and MAXDIV with feedback of 2 best users are also considered. It can be seen that MAXDIV-RF provides near-optimal peformance. Further, while MAXDIV-2best provides gains over DIV-OPT, MAXDIV-RF still provides about 35% gains over MAXDIV-2best. This can be attributed to the clever choice of the two feedback users in MAXDIV-RF that depends on the distributions of average and instantaneous user thoughputs, and hence does not necessarily map to the two best users.

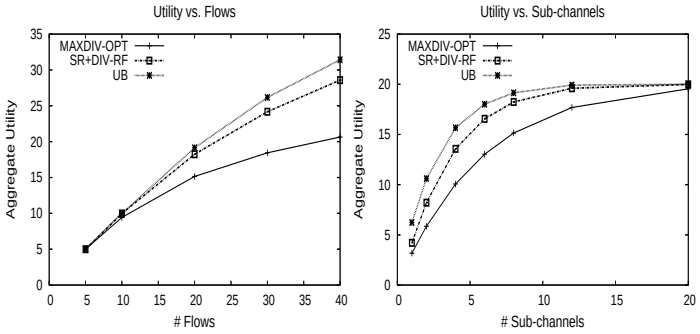
The feedback overhead needed for optimal performance, pre-

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938



a) 12 sub-channels, 3 relays b) 40 flows, 3 relays

Figure 7: QoS-RF performance in bottleneck and general scenarios.



a) Utility vs. Users: (x,8C,5R) b) Utility vs. Channels: (20F,x,5R)

Figure 8: Exploiting diversity and spatial reuse.

sented in Figures 6(a) and (b), indicate that the overhead increases very slowly with increasing users and channels and saturates. This is because, the number of feedback elements required has been shown to depend only on the average throughput deviation between users and the instantaneous rate deviation across channels and not on the specific number of users, channels and relays. The results convincingly establish and confirm that the number of users' feedback required for optimal performance is indeed small (≈ 2), explaining the near-optimal performance of MAXDIV-RF.

Figure 7 compares performance of QoS-RF against an upper bound for both bottleneck (relay hop) and general cases. Since MAXDIV with full feedback optimally solves the problem with no QoS constraints, it serves as a loose upper bound for the QoS version. It can be clearly seen that average case performance in bottleneck case is much closer to upper bound than the guarantee of half. The performance in general case is also good, being within a factor of half.

5.2 Interference-aware RS

We now evaluate the performance of SR+DIV-RF against that of optimal MAXDIV (full feedback) 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 effective rates of the scheduled users to correspond directly to their relay link rates in the absence of interference. The aggregate utility results for SR+DIV-RF are presented as a function of increasing users (flows) and sub-channels in Figures 8(a) and (b) respectively. It can be seen that for a given network capacity, when the number of users is small, the injected traffic load can be sustained even without exploiting spatial reuse. However, when the number of users increases, spatial reuse must be exploited to sustain a larger fraction of the injected traffic load. This in turn results in the gain of SR+DIV-RF increasing to 50% over MAXDIV-RF. These utility results indicate the superiority of SR+DIV-RF not just in utilization but also in fairness.

In addition to the significant gains over MAXDIV-RF, SR+DIV-RF 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 and is not worth the additional feedback required that would scale with the number of users.

6. CONCLUSIONS

Given the emerging applications of relay-enabled cellular networks, we have focused on the design of efficient scheduling algorithms for such networks in this work. The proposed solutions leverage the additional diversity and spatial reuse gains provided by these networks efficiently in trying to obtain the capacity around the base station, while also ensuring proportional fairness amongst the users. The presence of relays is uniquely exploited in the scheduling algorithms to effectively leverage diversity and spatial reuse gains. More importantly, all the proposed schedulers incur a scalable feedback overhead that grows only with the number of relays and not with the number of users unlike conventional approaches.

7. REFERENCES

[1] A. So and B. Liang, "Effect of relaying on capacity improvement in wireless local area networks," in *IEEE WCNC*, Mar 2005.

[2] P. Herhold, W. Rave, and G. Fettweis, "Relaying in cdma networks: pathloss reduction and transmit power savings," in *IEEE VTC*, Apr 2003.

[3] 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.

[4] Z. Zhang, Y. He, and K. P. Chong, "Oppotunistic downlink scheduling for multiuser ofdm systems," in *IEEE WCNC*, Mar 2005.

[5] G. Brar, D. Blough, and P. Santi, "Computationally efficient scheduling with the physical interference model for throughput improvement in wmns," in *ACM MOBICOM*, Sep 2006.

[6] D. Comstockand J. Lee, S. Zheng, and A. Zhang, "A flexible multi-hop frame structure for IEEE 802.16j," *IEEE 802.16 Broadband Wireless Access Working Group*, Nov 2006.

[7] N. Challa and H. Cam, "Cost-aware downlink scheduling of shared channels for cellular networks with relays," in *IEEE ICPCC*, 2004.

[8] 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.

[9] M. Herdin, "A chunk based ofdm amplify-and-forward relaying scheme for 4g mobile radio systems," in *IEEE ICC*, Jun 2006.

[10] A. Hottinen and T. Heikkinen, "Subchannel assignment in ofdm relay nodes," in *Proc. of CISS*, Mar 2006.

[11] K. Sundaresan and S. Rangarajan, "On exploiting diversity and spatial reuse in relay-enabled wireless networks," in *ACM MOBIHOC*, May 2008.

[12] J. Jang and K. B. Lee, "Transmit power adaptation for multi-user OFDM systems," *IEEE JSAC*, vol. 21, no. 2, pp. 171–179, 2003.

[13] B. Radunovic and J. Le Boudec, "Rate performance objectives of multi-hop wireless networks," in *IEEE INFOCOM*, Mar 2004.

[14] C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity (Chapter 11)*, Prentice Hall, 1982.

[15] M. R. Garey and D. S. Johnson, *Computers and intractability: A guide to the theory of NP-completeness*, W. H. Freeman and Company, 1979.

[16] R. G. Mukthar, "Qns: Queuing network simulator," in *QNS v0.1*, http://www.cubinlab.ee.mu.oz.au/rgmukht/qns, Nov 2003.

Digital Object Identifier: 10.4108/ICST.WICON2008.4938
http://dx.doi.org/10.4108/ICST.WICON2008.4938