Exploiting Multi-user Diversity and Multi-hop Diversity in Dual-hop Broadcast Channels

Ammar Zafar, *Student Member, IEEE*, Mohammad Shaqfeh, *Member, IEEE*, Mohamed-Slim Alouini, *Fellow, IEEE*, and Hussein Alnuweiri, *Member, IEEE*

***Abstract*—We propose joint user-and-hop scheduling over dual-hop block-fading broadcast channels in order to exploit multi-user diversity gains and multi-hop diversity gains all together. To achieve this objective, the first and second hops are scheduled opportunistically based on the channel state information. The joint scheduling problem is formulated as maximizing the weighted sum of the long term achievable rates of the users under a stability constraint, which means that in the long term the rate received by the relay should equal the rate transmitted by it, in addition to power constraints. We show that this problem is equivalent to a single-hop broadcast channel by treating the source as a virtual user with an optimal weight that maintains the stability constraint. We show how to obtain the source weight either off-line based on channel statistics or on real-time based on channel measurements. Furthermore, we consider special cases including the maximum sum-rate scheduler and the proportional fair scheduler. We also show how to extend the scheme into one that allows multiple user scheduling via superposition coding with successive decoding. Numerical results demonstrate that our proposed joint scheduling scheme enlarges the rate region as compared to scheduling schemes that exploit the diversity gains partially.**

***Index Terms*—Broadcast channels, dual-hop, block-fading, multi-hop diversity, multi-user diversity, joint scheduling, optimal resource allocation.**

I. Introduction

The classical relay channel model was first proposed in the information theory literature in the late 1960's and early 70's [2]–[4]. However, the practical use of relaying in wireless communication systems has been taken into consideration in the last few years. The recent advances in wireless communications and the growing demands for wireless connectivity have rekindled interest in relays. Consequently, much research work has been carried out on relaying systems recently and it has been shown that relaying can enhance the coverage and capacity of wireless networks. In particular, relaying can evidently enhance the transmission capacity for users at the edge of a wireless cell [5]–[10]. Due to these advantages, relays are now stated in telecommunications standards [11].

In the Long Term Evolution (LTE)-Advanced systems, fixed access point relays with only an in-band wireless connection to the backhaul network are to be deployed [12]–[14]. One of the major challenges for LTE-Advanced systems is to utilize the precious air-link (i.e. bandwidth) resources efficiently in order to enable achieving the prospected high Quality-of-Service (QoS) requirements [15]. Therefore, the resource allocation task is of crucial importance and we aim to investigate this problem taking into consideration the use of relays in LTE-Advanced systems.

Two relay functionalities are considered in LTE-Advanced [11]; (i) Transparent relays in which a user has a direct link with the base station in addition to the relay link, and (ii) Non-transparent relays which extend the coverage of a cell to reach remote users with no connection to the base station. The latter follows a dual-hop channel model [16]. In [17], optimal joint power and resource allocation for the transparent relays case was considered. The problem was formulated by maximizing the achievable rate region of block-fading relay-assisted broadcast channels (i.e. for the downlink). The relays were assumed to operate using decode-and-forward (DF). Closed-form formulas were provided to decide when to seek relay support and how to integrate this with the opportunistic multi-user scheduling task. In [18], the uplink case was discussed and optimal resource allocation to maximize the achievable rate region of relay-assisted multiple access channels was obtained. However, reference [18] considered an additive white Gaussian noise (AWGN) channel with a constant channel gain. There are other works considering resource allocation strategies for relay networks from different prospectives such as [19]–[22]. In [19] and [20] the single relay and single user channel was considered from information theoretic perspective. In [21], [22] and [23] resource allocation with user cooperation was considered. However, in this work we consider the case of fixed access point relays which are not users' terminals.

In this paper, we consider the downlink of a multi-user network with the users connected to the source through a non-transparent relay under block-fading channels. Hence, there are no direct connections between the source and the users. Furthermore, we consider decode-and forward relaying and half-duplex relays as described in more detail in Sec-

A. Zafar and M.-S. Alouini are with the Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Building 1 Level 3, King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia. 23955-6900 (E-mail: {ammar.zafar, slim.alouini}@kaust.edu.sa).

M. Shaqfeh and H. Alnuweiri are with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, C/o Qatar Foundation, PO Box 23874, Doha, Qatar. (E-mail: {Mohammad.Shaqfeh, Hussein.Alnuweiri}@qatar.tamu.edu).

This paper was made possible by YSREP grant # 2-011-2-002 from the Qatar National Research Fund (a member of Qatar Foundation). Furthermore, KAUST funded the efforts of A. Zafar partially and the efforts of M.-S. Alouini. The statements made herein are solely the responsibility of the authors.

This work was partially presented in IEEE Broadband Wireless Access Workshop (BWA 2012), Anaheim, CA [1].

tion II. Our contribution is providing the optimal resource allocation scheme by exploiting two important sources of gains in the system which are **multi-user diversity (MUD)** [24] and **multi-hop diversity (MHD)** [25] jointly. As well-known in the literature [26], multi-user diversity is due to the independent fading conditions of the users' channels in a wireless network, which make it more likely that at each instance there is one user with very good channel condition. Hence, MUD can be obtained by using *channel-aware user scheduling* to allocate the wireless resources to the users dynamically based on their instantaneous channel conditions. A better way to maximize MUD gains is to allow multiple users to be scheduled simultaneously based on the channel conditions by applying *superposition coding with successive interference cancellation for decoding* at the receiver, which enables achieving the capacity of degraded broadcast channels as we know from the information theory literature [27], [28]. Similarly, multi-hop diversity is due to the independent fading conditions over the two hops of communication. Therefore, MHD can be obtained by using *channel-aware hop scheduling* by selecting to transmit over the hop that is in a good state. As a prerequisite, the relay should have the ability to store the received messages from the first hop in buffers instead of forwarding them over the second hop directly without taking the channel state into consideration. We refer to this as a *buffering gain* (Buff), which is an elementary component to obtain full MHD gains. To improve the Buff gain into full MHD gain, the relay should forward the source messages over the second hop when it has good channel condition. In this work, we investigate these sources of diversity gains and we compare the performance of scheduling schemes that exploit them partially versus an optimal scheme that exploits both MUD and MHD all together.

Our main contribution is proposing a novel joint user-and-hop scheduling scheme that opportunistically allocates the resources based on instantaneous channel measurements. We provide rigorous discussion about the proposed scheduler. We discuss both single user scheduling and superposition coding with successive interference cancellation for the joint scheduler. Furthermore, we show how to optimize the optimal scheduling scheme off-line based on channel statistics or on real-time based on channel measurements. Moreover, we provide closed-form formulas to characterize the long-term average achievable rates via the proposed scheduler. Also, we discuss the scheduling criterion in case of variable power in addition to the constant power allocation case. Furthermore, we consider two common special cases by applying the joint scheduling scheme for (i) the maximum sum throughput scheduler and (ii) the proportional fair scheduler. On top of that, in order to characterize the obtained MHD gains, we compare the performance versus two benchmarks with no MHD gains. Specifically, we compare the proposed joint scheduler with two conventional scheduling schemes; (i) The first one applies multi-user scheduling alone yielding a MUD gain only, and (ii) the second one is a round robin scheduler [29], [30] which does not incorporate channel state in the resource allocation and hence it does not yield neither MUD nor MHD gains. Additionally, in order to have a comprehensive analysis, we quantify also the gain achieved by buffering at the relay (Buff gain) without achieving the full MHD gains for these two conventional schedulers that are used as benchmarks for the comparison with the optimal scheme. We provide thorough numerical results to demonstrate our results and findings.

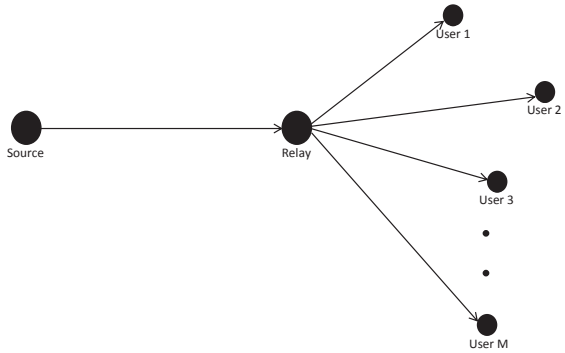

Fig. 1: System model: dual-hop broadcast channel.

The rest of the paper is organized as follows. Section II describes the system model and the problem formulation. In Section III, we discuss a simple resource allocation scheme with neither hop nor user scheduling, and in Section IV, we discuss a conventional scheduler with multi-user scheduling only. Next, we discuss the proposed joint hop-and-user scheduling scheme in Section V. Then, numerical results are presented in Section VI. After that we provide in Section VII analysis of the probability of getting an empty buffer at the relay. Finally, Section VIII concludes the paper.

II. System Model

We consider a system with a source (S), a relay (R) and M users or destinations (D). There are no direct links between the source and the M users. Hence, the source transmits to the users only through the relay as shown in Fig. 1. A block-fading channel model is assumed. One channel block, called a resource unit (RU), can consist of multiple time slots and multiple frequency sub-carriers. However, the RUs are orthogonal to each other (i.e. they do not overlap over time or frequency). The relay is assumed to work in half-duplex mode, which means that the relay cannot receive and transmit simultaneously over the same RU. Furthermore, the relay follows the decode-and-forward (DF) protocol. The fading gains of the S-R link and the R-D links stay constant over a single RU. However, they change randomly and independently from one RU to another. Moreover, the channel gains of all links, i.e. S-R and R-D, experience independent fading. Without loss of generality we assume that all RUs have the same duration and bandwidth. We also assume that the relay transmits to only one user within one RU. This assumption is omitted in Section V-D when we discuss multiple user scheduling per RU via superposition coding (SC). Furthermore, for some of the schemes that are introduced in this paper, we assume that

either the source or the relay can transmit in one RU and not both of them. On the other hand, for some other schemes, we assume that every RU is divided orthogonally in the time domain into two sub-blocks such that the source transmits in the first sub-block and then the relay transmits in the second sub-block of the same RU.

The relay has channel state information (CSI) of its channels with all users and with the source[1]. The objective is to exploit the channel knowledge in the resource allocation task in order to maximize the achievable rates by obtaining MUD and MHD gains. To achieve this target, we propose to apply *joint user-and-hop scheduling (JS) scheme* in which the first and second hops are scheduled opportunistically based on CSI. We assume that the relay has the ability to store the received messages from the source in buffers until the channel of the destined user becomes good and this user gets scheduled. Furthermore, we assume that when the source is scheduled, it may combine the packets of more than one user together within one RU, and the relay distributes the source message across the data buffers for each user. The source and the relay adjust the transmission rates based on the instantaneous channel condition (capacity). Moreover, we assume best effort data traffic so that there are no delay constraints and the objective function is to maximize the weighted sum of the long-term average achievable rates (throughputs).

In the numerical results provided in Section VI, we assume that all nodes are equipped with a single antenna. The achievable rate (in bits/sec/Hz) within one RU, given index k, for the channel from the relay to the ith user, where the complete RU is allocated to this channel, is given by[2]

$$R_i^{ac}[k] = \log_2\left(1 + \frac{P[k]|h_i[k]|^2}{N_0}\right), \tag{1}$$

where $h_i[k]$ is the complex channel gain between the relay and the ith user in the kth RU ($i = 0$ is used to denote the source), $P[k]$ is the power in (Joul/sec/Hz) and N_0 is the noise power spectral density. For simplicity of notation, we use a single index k to denote the RUs although they span the time and frequency domains.

We formulate the resource allocation optimization problem as maximizing the weighted sum average[3] throughput of the M users. We need to apply a *stability constraint* meaning that the sum throughput to the users cannot exceed the throughput generated by the source. Moreover, we consider two cases in terms of power allocation. For constant power allocation we have $P[k] = \bar{P}$, while for optimal power allocation we have a maximum average (over all RUs) sum (for source and relay) power constraint $\bar{P}$. The main optimization problem (with both stability and power constraints) can be written as

$$\begin{aligned} &\max \quad \sum_{i=1}^{M} \mu_i \bar{R}_i, \quad \text{subject to} \\ &\bar{R}_0 = \sum_{i=1}^{M} \bar{R}_i, \quad \text{and} \quad \frac{1}{K}\sum_{k=1}^{K} P[k] \leq \bar{P}, \end{aligned} \tag{2}$$

where we give the S-R channel the index zero, μ_i is the specified $i$the user weight and $\bar{R}_0$ and $\bar{R}_i$ are the average throughput of the source and ith user, respectively. It should be clear that the optimization variables of problem (2) are the resource allocation decisions over the RUs which affect the average throughputs of the users. The characterization of the average throughput as a function of the scheduling criteria are derived throughout the following sections for the different schemes that we consider.

Although the constraint ($\bar{R}_0 = \sum_{i=1}^{M} \bar{R}_i$) in (2) guarantees the stability of the relays buffers because, on the long-term, the information rate received by the relay equals the information rate transmitted by it, there is a possibility that, due to the dynamic scheduling process, we may have instances at which the relay's buffers get empty and hence the relay can not be scheduled in this case. So, the source should be scheduled in this case regardless of its channel condition. Therefore, the obtained results for the achievable throughput are actually upper bounds which may be little bit degraded due to the empty buffers scenario. However, as an easy and practical solution to eliminate this case, we could assume, for example, that there is an initial state of source transmission only such that the relay's buffer gets occupied with sufficient amount of data bits. Then, we start applying the optimal opportunistic scheduling described in Section V. In this case, the probability of having the relay scheduled over many successive RUs until the buffers get empty becomes very negligible. To elaborate more on this point, we provide more comprehensive analysis of the empty buffer case in Section VII.

We provide the solution of (2) for generic users' weights μ, and we emphasize two special cases[4] which are maximum sum throughput (all μ's are equal) and proportional fairness ($\mu_i = 1/\bar{R}_i$) [26]. For now, we will neglect the power constraint and use $P[k] = \bar{P}$. The power constraint will be considered in Section V-B.

As benchmarks to be used for comparison with JS, we consider two scheduling schemes. The first scheme, named round robin scheduling, does not employ neither multi-user scheduling nor multi-hop scheduling. This scheme allocates part of the total bandwidth to every user without taking the instantaneous channel conditions into consideration. The second scheme employs multi-user scheduling only. For both schemes, we consider the scenario that buffering is not feasible at the relay and the other scenario is when buffering is possible at the relay.

[1]Full CSI is required for the scheduling schemes. If the scheduling decisions are done at the relay or at the backhaul network, the source does not necessarily require CSI of the users' channels in the second hop. Nevertheless, when the source is scheduled to transmit, it must know the channel capacity of its channel with the relay in order to adjust its transmission rate accordingly.

[2]This is not valid in case of SC which is discussed in Section V-D.

[3]We use average and long term interchangeably. Similarly we use throughput and rate.

[4]Maximizing a weighted sum of the rates guarantees Pareto-optimality, while the specific selection of the weights enables controlling the trade-off between throughput and fairness [31], [32].

III. Round Robin Scheduling

A. Without Buffering at the Relay

Round robin scheduling (RRS) with no buffering at the relay has no diversity gains since it does not exploit neither MUD nor MHD. Therefore, it gives us a lower bound on the performance of the system. Since the buffering capability is not available at the relay, it has to forward the message from the source to the intended user immediately after reception. Hence, each RU is divided into two sub-blocks in the time domain. In the first sub-block, the source transmits to the relay and in the second sub-block the relay transmits to the intended user. The duration of each sub-block is set such that the throughput of the S-R link and R-D link are equal, so that we can transmit at channel capacity (1) while the stability constraint on the system is satisfied. An example of allocation of RUs for this scheme is shown in Fig. 2(a). The throughput of the source and the ith user in the kth RU is

$$R_0[k] = \tau_{si}[k] R_0^{ac}[k], \qquad R_i[k] = \tau_{ri}[k] R_i^{ac}[k], \tag{3}$$

where $\tau_{si}[k]$ is the ratio of the duration of the sub-block allocated to the first hop to the total duration of kth RU, and $\tau_{ri}[k]$ is similarly the ratio of the sub-block allocated to the second hop to the total duration of kth RU of the ith user. Thus, $\tau_{si}[k] + \tau_{ri}[k] = 1$. To maximize the achievable rate while maintaining the stability constraint $\tau_{si}[k]$ and $\tau_{ri}[k]$ are adjusted such that $R_0[k]$ and $R_i[k]$ are equal. Therefore, we have $\tau_{si}[k] = \frac{R_i^{ac}[k]}{R_0^{ac}[k]+R_i^{ac}[k]}$ and $\tau_{ri}[k] = \frac{R_0^{ac}[k]}{R_0^{ac}[k]+R_i^{ac}[k]}$.

The average throughput of the i^{th} user is then given by

$$\bar{R}_i = \tau_i \int_0^\infty \int_0^\infty \frac{r_0 r_i}{r_0 + r_i} f_{R_0}(r_0) f_{R_i}(r_i) dr_0 dr_i, \tag{4}$$

where τ_i is the ratio of blocks in which the i^{th} user is scheduled, $f_R(r)$[5] is the probability density function (PDF) of the achievable rate, R^{ac}, and $F_R(r)$ is the cumulative distribution function (CDF) of it. The channel ratio τ_i can be set according to the needs of each user[6].

Note that the average total throughput, which is equivalent to the average throughput transmitted by the source, $\bar{R}_0$, can be obtained by summing the average throughput of all users.

B. With Buffering at the Relay

In this subsection, we consider the round robin scheme with buffering capability available at the relay (RRS-Buff). Hence, the system can now benefit from Buff gain. Now, the source and relay do not need to transmit in the same RU. Instead, the source is scheduled for a certain number of RUs and the rest of the RUs are used for the relay to serve the users as shown in Fig. 2(b). Thus, a portion of the bandwidth is reserved for the source and for each user respectively. The ratios of the total bandwidth that are allocated to each link are adjusted based on the needs of the users and by taking into consideration maintaining the stability constraint of the system. The relay stores the received messages from the source and transmits them to the respective users over their reserved channels. We denote the ratio of RUs in which the source is scheduled as τ_s and the ratio of RUs in which the relay transmits to the users as τ_R ($\tau_s + \tau_R = 1$). Furthermore, τ_R is sub-divided into M blocks, each with ratio of τ_i, where $\sum_{i=1}^{M} \tau_i = 1$. Therefore, the average achievable average rate of the source and the ith user are given by

$$\bar{R}_0^{ac} = \int_0^\infty r f_{R_0}(r) dr \tag{5}$$

$$\bar{R}_i^{ac} = \tau_i \int_0^\infty r f_{R_i}(r) dr \tag{6}$$

The long term throughput of the source and the ith user can then be obtained from their respective achievable rates as

$$\bar{R}_0 = \tau_s \bar{R}_0^{ac} \quad \text{and} \quad \bar{R}_i = \tau_R \bar{R}_i^{ac}, \tag{7}$$

From the stability constraint in (2) and noting that $\tau_s + \tau_R =$

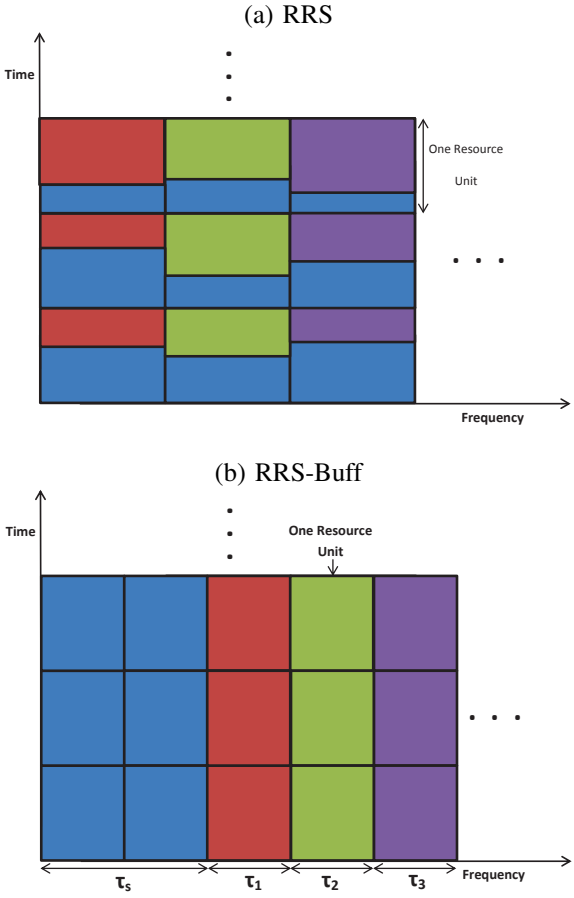

Fig. 2: Example of the allocation of RUs for (a) RRS (no gains) (b) RRS-Buff (Buff gain). Blue represents the source (first hop), while red, green, and purple represent three users.

[5]For Rayleigh fading, $f_R(r) = \frac{\ln 2}{\rho} 2^r \exp\left(-\frac{2^r-1}{\rho}\right) \quad r \geq 0$, and $F_R(r) = 1 - \exp\left(-\frac{2^r-1}{\rho}\right) \quad r \geq 0$, where ρ is the SNR [33].

[6]For instance, in a two user system, one user might be requesting video and the other audio. Hence, the user which requests the video will have a greater τ_i as it requires higher data rate.

1, we obtain that τ_s and τ_R should be adjusted according to

$$\tau_s = \frac{\sum_{i=1}^{M} \bar{R}_i^{ac}}{\sum_{i=1}^{M} \bar{R}_i^{ac} + \bar{R}_0^{ac}}, \quad \tau_R = \frac{\bar{R}_0^{ac}}{\sum_{i=1}^{M} \bar{R}_i^{ac} + \bar{R}_0^{ac}}. \tag{8}$$

IV. Multi-user Scheduling Only

In this section, we consider the MUD gain achieved by multi-user scheduling only. Similar to the previous section, we consider when buffering is not available at the relay and the alternative case when buffering is possible at the relay. For each scenario, we discuss scheduling in the general case of having a given weight factor for each user as well as the particular case of proportional fair scheduling.

A. Without Buffering at the Relay

1) Conventional Scheme with given Weights: In this scheme, the users are opportunistically scheduled in each RU according to their respective channel conditions in that RU. Thus, MUD gain is achieved because users are scheduled when their channels are in good states. However, the system does not exploit MHD since we do not have hop scheduling. Furthermore, similar to the scheme in Section III-A, since the relay does not store the source messages in a buffer, every RU is divided into two sub-blocks of ratios τ_{si} and τ_{ri} for the source and the relay respectively, as shown in Fig. 3(a). Furthermore, τ_{si} and τ_{ri} are set such that the rates of the source and the scheduled user in the RU are equal. Therefore, if the ith user is scheduled in the kth RU, the transmission rate will be

$$R_i[k] = \tau_{ri}[k] R_i^{ac}[k] = \frac{R_0^{ac}[k] R_i^{ac}[k]}{R_0^{ac}[k] + R_i^{ac}[k]}, \tag{9}$$

where $R_i^{ac}[k]$ and $R_0^{ac}[k]$ follow (1). Therefore, the long term throughput of the ith user is given by

$$\bar{R}_i = \int_0^\infty \int_0^\infty \frac{r_0 r_i}{r_0 + r_i} f_{R_0}(r_0) f_{R_i}(r_i) \times \text{Prob}(i\text{th user is scheduled} | R_i^{ac} = r_i, R_0^{ac} = r_0) dr_0 dr_i, \tag{10}$$

where $\text{Prob}(i\text{th user is scheduled} | R_i^{ac} = r_i, R_0^{ac} = r_0)$ is the conditional probability that the ith user is selected in a RU given that the ith user achievable rate in that RU is $R_i^{ac} = r_i$ and the source achievable rate $R_0^{ac} = r_0$. Notice that the expression (10) is similar to (4) except for the additional probability term and the removal of the resources ratio term τ_i. This is because, unlike the scheme in Section III-A, we do not have, in this case, certain ratios of RUs reserved for every user. The user is allocated a RU based on its achievable rates in comparison with the other users. Hence, the user selection criterion is such that the mth user is selected according to

$$m = \arg \max_i \mu_i R_i[k] \quad i = 1, \ldots, M, \tag{11}$$

where $R_i[k]$ is given by (9). From (11) user i is selected when

$$\mu_l \frac{R_0^{ac}[k] R_l^{ac}[k]}{R_0^{ac}[k] + R_l^{ac}[k]} < \mu_i \frac{R_0^{ac}[k] R_i^{ac}[k]}{R_0^{ac}[k] + R_i^{ac}[k]}, \quad \forall\, l \neq i. \tag{12}$$

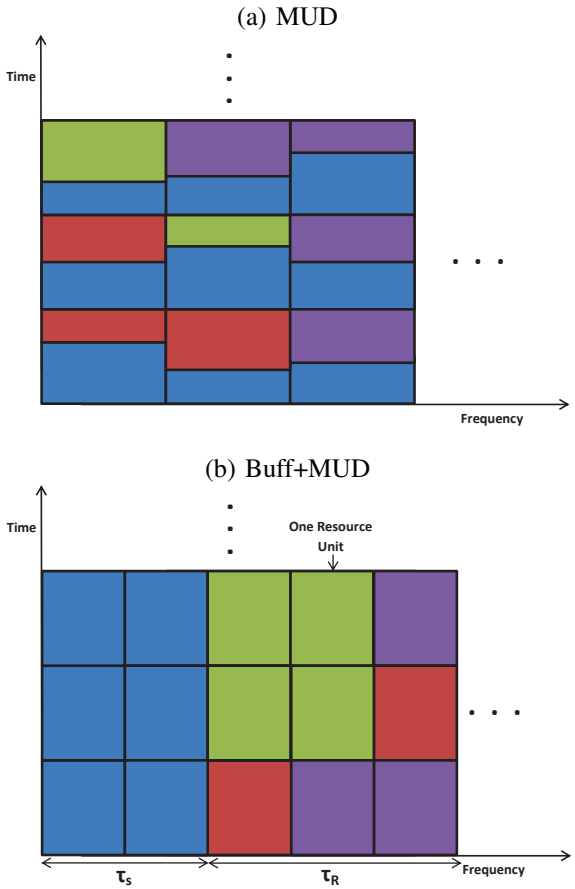

Fig. 3: Example of the allocation of RUs for multi-user scheduling only (a) without buffering (MUD) (b) with buffering (Buff+MUD). Blue represents the source (first hop), while red, green, and purple represent three users.

With some manipulations, we can write (12) as (where we remove the RU index k, and replace $R_i^{ac}[k]$, $R_l^{ac}[k]$, $R_0^{ac}[k]$ by r_i, r_l, r_0, respectively, for simplicity of notation)

$$\frac{\mu_l}{\mu_i r_i} + \frac{\mu_l}{\mu_i r_0} - \frac{1}{r_0} < \frac{1}{r_l}, \quad \forall\, l \neq i. \tag{13}$$

Note that the left hand side of (13) can in principle be positive or negative depending on the weighting factors μ_i and μ_l. Therefore, we can distinguish two cases for the conditional probability that the ith user is selected.

$$\text{Prob}\left(\frac{\mu_l}{\mu_i r_i} + \frac{\mu_l}{\mu_i r_0} - \frac{1}{r_0} < \frac{1}{r_l}\right) = \begin{cases} 1 & : \frac{\mu_l}{\mu_i r_i} + \frac{\mu_l}{\mu_i r_0} - \frac{1}{r_0} < 0 \\ F_{R_l}\left(\frac{\mu_i r_i r_0}{\mu_l r_0 + \mu_l r_i - \mu_i r_i}\right) & : \frac{\mu_l}{\mu_i r_i} + \frac{\mu_l}{\mu_i r_0} - \frac{1}{r_0} \geq 0. \end{cases} \tag{14}$$

We apply a similar notation to the one suggested in [34] to write the conditional probability in a compact form as

$$\text{Prob}(i\text{th user is scheduled} | R_i = r_i, R_0 = r_0) = \prod_{l \neq i} F_{R_l}\left(\left[\frac{\mu_i r_i r_0}{\mu_l r_0 + \mu_l r_i - \mu_i r_i}\right]^*\right), \tag{15}$$

where

$$[y]^* = \begin{cases} y & y \geq 0 \\ \infty & y < 0. \end{cases} \quad (16)$$

By substituting (15) in (10) we characterize the long term throughput of the ith user. The long term source throughput is obtained by summing the throughputs of all users.

2) Proportional Fair Scheduling: To achieve proportional fairness (PF) [26], the users' weights are given by $\mu_i = \frac{1}{\bar{R}_i}$. To obtain the optimal weights off-line, we substitute for $\bar{R}_i$ using (10) and hence we get a total of M equations and M unknown μ_i's. However, they are difficult to solve analytically. Therefore, we propose a recursive algorithm to obtain the optimal μ_i's by simulating a real-time implementation of the PF scheduler. The update equation for μ_i is given by

$$\mu_i[n] = \frac{1}{\bar{R}_i[n] + d[n]}, \quad (17)$$

where n represents the iteration number (which is equivalent to the time index in real-time implementation), $d[n] = 1 - \frac{n}{1000}$ and $\bar{R}_i[n]$ is the average rate of user i for all iterations up to n. The function $d[n]$ is used because average throughputs of all the users are initialized to zero. $d[n]$ is set to zero after the thousandth (selected arbitrary) iteration. The μ_i's are updated in each iteration and these updated μ_i's are then used to schedule the users in the next iteration and so on until we converge to the steady state value for the optimal weights for PF. The complete algorithm is shown in Algorithm 1.

Algorithm 1 Proportional Fair-(BMUD)

Initialize $\bar{R}_i = 0 \ \ \forall\ i = 1, \cdots, M$
for $n = 1$ to N **do**
  Generate the channel gains randomly based on the fading channel model
  $d = 1 - \frac{n}{1000}$
  if $d < 0$ **then**
    $d = 0$
  end if
  $\mu_i(n) = \frac{1}{\bar{R}_i + d} \ \ \forall\ i$
  R_i follows (9)
  $[R^*, ind] = \max(\mu_i R_i)$
  $R_{ind} = R^*$
  $R_k = 0 \ \ \forall\ k \neq ind$
  $\bar{R}_i(n+1) = \bar{R}_i(n) + \frac{1}{n+1}\left(R_i - \bar{R}_i(n)\right) \ \ \forall\ i$
end for
return $\mu_i(N) \ \ \forall\ i$

B. With Buffering at the Relay

1) Conventional Scheme with given Weights: In this scheme (refer to Fig. 3(b)), we can obtain MUD gains due to multi-user scheduling and a buffering gain which is obtained because the relay does not have to immediately forward the data from the source to the intended user. Instead of that, it stores the data in its buffers and forwards it later when the user's channel is good. However, we do not obtain full MHD gains because a certain portion of the RUs are allocated to the S-R link regardless of the instantaneous channel values of the S-R link. The ratio of RUs where the source is scheduled is denoted τ_s, while the ratio of RUs in which the relay transmits to the users is denoted τ_R. The selection of τ_s and τ_R is based on maintaining the stability constraint. The RUs in τ_R are allocated opportunistically by selecting the mth user to be scheduled according to (for constant power allocation)

$$m = \arg\ \max_i \mu_i R_i[k] \ \ i = 1, \ldots, M, \quad (18)$$

where $R_i[k]$ follows (1). The average achievable rate by the source and ith user can be written as $\bar{R}_0 = \tau_s \bar{R}_0^{ac}$ and $\bar{R}_i = \tau_R \bar{R}_i^{ac}$ respectively, where

$$\bar{R}_0^{ac} = \int_0^\infty r f_{R_0}(r) dr \quad (19)$$

$$\bar{R}_i^{ac} = \int_0^\infty r f_{R_i}(r) \text{Prob}(i\text{th user is scheduled} | R_i = r) dr, \quad (20)$$

$$\text{Prob}(i\text{th user is scheduled} | R_i = r) = \prod_{l \neq i} F_{R_l}\left(\frac{\mu_i r}{\mu_l}\right). \quad (21)$$

τ_s and τ_R are adjusted such that the stability constraint is maintained. Therefore, we have $\tau_s + \tau_R = 1$ and $\tau_s \bar{R}_0^{ac} = \sum_{i=1}^M \tau_R \bar{R}_i^{ac}$, which yields

$$\tau_s = \frac{\sum_{i=1}^M \bar{R}_i^{ac}}{\sum_{i=1}^M \bar{R}_i^{ac} + \bar{R}_0^{ac}}, \quad \tau_R = \frac{\bar{R}_0^{ac}}{\sum_{i=1}^M \bar{R}_i^{ac} + \bar{R}_0^{ac}}. \quad (22)$$

2) Proportional Fair Scheduling: We again consider the special case of PF scheduling. As discussed previously, the users' weights are given by $\mu_i = \frac{1}{\bar{R}_i}$, which is in this case also equivalent to $\mu_i = \frac{1}{\bar{R}_i^{ac}}$ since the ratios of the weights is the same in both cases. Hence, the algorithm for finding the weights is the same as in Algorithm 1 except that R_i is computed using (1) instead of (9).

V. Joint User-and-Hop Scheduling

A. Single User Selection per Resource Unit

In Section IV-B, the source and the users were scheduled independently. Hence, multi-hop diversity was not exploited properly. In this section, we show that the optimization problem in (2) can be maximized by joint scheduling the source and the M users. By converting the constrained optimization problem in (2) into an unconstrained problem using the Lagrangian dual problem, we can write

$$\begin{aligned} \max \ \ & \sum_{i=1}^M \mu_i \frac{1}{K} \sum_{k=1}^K R_i[k] - \frac{\mu_0}{K}\left(\sum_{i=1}^M \sum_{k=1}^K R_i[k] - \sum_{k=1}^K R_0[k]\right) \\ &= \frac{1}{K} \sum_{k=1}^K \ \max \ \ \sum_{i=1}^M (\mu_i - \mu_0) R_i[k] + \mu_0 R_0[k], \end{aligned} \quad (23)$$

where we have assumed constant power per block (optimal power allocation is considered in the next subsection). We

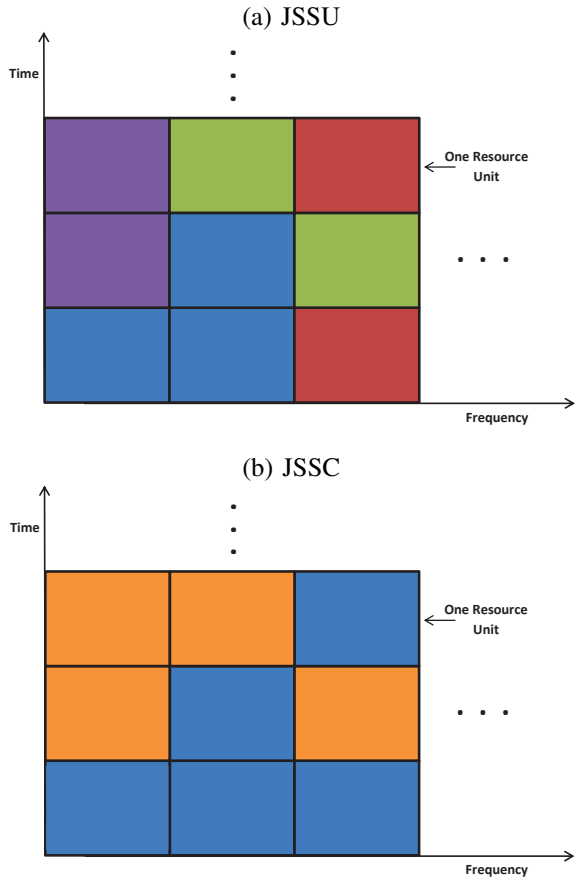

Fig. 4: Example of the allocation of RUs for joint user-and-hop scheduling (a) single user selection per RU (JSSU) (b) superposition coding with successive interference cancellation (JSSC). Blue represents the source (first hop), while red, green, and purple represent three users and yellow represents the combination of the three users.

observe that the Lagrangian dual variable μ_0 for the stability constraint appears as a weighting factor for the source. Furthermore, to maximize the weighted sum of the rates on the long-term, the optimal solution involves maximizing the weighted sum of the achievable rates for each RU independently. We can see from (23) that maximizing the initial problem in (2) is equivalent to, assuming half-duplex relaying and single user selection per block, selecting the source or the user with the maximum weighted instantaneous rate in each RU where the source weight, μ_0, is optimized to maintain the stability constraint in (2). Therefore, the system can be thought of having a single-hop broadcast channel with $M+1$ users with the source acting as a virtual user and the relay acting as the virtual source. The selection criteria can be written as, where we use index 0 for the source,

$$m = \arg \max_j \mu_j^{'} R_j[k] \quad j = 0, 1, \ldots, M, \tag{24}$$

where

$$\mu_j^{'} = \begin{cases} \mu_0 & j = 0 \\ (\mu_j - \mu_0)^+ & j = 1, 2 \ldots, M, \end{cases} \tag{25}$$

where $(x)^+ = \max(x, 0)$. Therefore, either the source or one of the users is scheduled in each RU depending upon (24) as shown in Fig. 4(a). So, unlike the scheme in Section IV-B, there is no specific ratio of blocks in which the source is scheduled. It depends merely on the instantaneous channel conditions.

The long term achievable rates are given by

$$\bar{R}_0 = \int_0^\infty r f_{R_0}(r) \text{Prob}(\text{source is scheduled} | R_0 = r) dr \tag{26}$$

$$\bar{R}_i = \int_0^\infty r f_{R_i}(r) \text{Prob}(i\text{th user is scheduled} | R_i = r) dr, \tag{27}$$

From (24), we can obtain

$$\text{Prob}(\text{source is scheduled} | R_0 = r) = \prod_{i=1}^{M} F_{R_i}\left(\frac{\mu_0 r}{\mu_i^{'}}\right) \tag{28}$$

$$\text{Prob}(i\text{th user is scheduled} | R_i = r) = F_{R_0}\left(\frac{\mu_i^{'} r}{\mu_0}\right) \prod_{l \neq i, 0} F_{R_l}\left(\frac{\mu_i^{'} r}{\mu_l^{'}}\right). \tag{29}$$

The optimal value of μ_0 can be obtained by solving $\bar{R}_0 = \sum_{i=1}^{M} \bar{R}_i$, where $\bar{R}_0$ and $\bar{R}_i$ are obtained using (26) and (27) respectively. This problem can be solved numerically using a one-dimensional bisection search over μ_0.

B. Optimal Power Allocation

Up till now, constant power per RU has been considered. We consider here the optimal power allocation. Due to the additional constraint, the equivalent dual problem is given by

$$\frac{1}{K} \sum_{k=1}^{K} \max \sum_{i=1}^{M} (\mu_i - \mu_0) R_i[k] + \mu_0 R_0[k] - \sum_{j=0}^{M} \frac{\lambda P_j[k]}{N_0}, \tag{30}$$

where λ is the dual variable which should be adjusted to maintain the average power constraint. Similar to the constant power case, the scheduling problem is equivalent to the conventional single-hop case by treating the source as a virtual user with weighting factor μ_0. The scheduling rule in the kth RU is now done according to

$$m = \arg \max_j \left(\mu_j^{'} R_j[k] - \frac{\lambda P_j[k]}{N_0}\right) \quad j = 0, 1, \ldots, M, \tag{31}$$

where $\mu_j^{'}$ is defined in (25), and $P_j[k]$ is given by a water-filling formula,

$$P_j[k] = \begin{cases} N_0 \left[\frac{\mu_j^{'}}{\lambda} - \frac{1}{|h_j[k]|^2}\right]^+ & j = m \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

To characterize the long-term rates, we use the PDF and CDF of the channel power gain, $|h|^2$, denoted respectively as $f_h(x)$ and $F_h(x)$[7] instead of $f_R(r)$ and $F_R(r)$ since the variable power allocation alters the statistics of the achievable

[7]For Rayleigh fading, $f_h(x) = \frac{1}{\bar{\gamma}} \exp\left(-\frac{x}{\bar{\gamma}}\right)$, $F_h(x) = 1 - \exp\left(-\frac{x}{\bar{\gamma}}\right)$, where $\bar{\gamma}$ is the average channel power gain.

rate. The single-user scheduling with optimal power allocation for conventional broadcast channel was studied in [35]. We apply the derived equations there to our case as follows. The power price parameter λ can be obtained from the total average power constraint given by

$$N_0 \sum_i \int_{\frac{\lambda}{\mu_i^{'}}}^{\infty} f_{h_i}(x) \prod_{j \neq i} F_{h_j}(\zeta) \left[\frac{\mu_i^{'}}{\lambda} - \frac{1}{x} \right] dx = \bar{P}, \quad (33)$$

which can be solved numerically using a bisection search, where ζ equals

$$\zeta = \frac{-\lambda}{\mu_j^{'} W \left[\left(-\frac{\lambda}{\mu_i^{'} x} \right)^{\frac{\mu_i^{'}}{\mu_j}} \exp\left(\frac{\mu_i^{'}}{\mu_j^{'}} - \frac{\lambda}{\mu_j^{'} x} - 1 \right) \right]}, \quad (34)$$

and $W[.]$ is the Lambert function. The long term achievable rates are given by

$$\bar{R}_0 = \int_0^{\infty} \log_2 \left(\frac{\mu_0 x}{\lambda} \right) f_{h_0}(x) \prod_{i=1}^{M} F_{h_i}(\zeta) dx \quad (35)$$

$$\bar{R}_i = \int_0^{\infty} \log_2 \left(\frac{\mu_i^{'} x}{\lambda} \right) f_{h_i}(x) \prod_{j \neq i}^{M} F_{h_j}(\zeta) dx. \quad (36)$$

Similar to the constant power case, μ_0 is obtained by equating the average source rate to the sum of the average users' rates.

C. Real Time Adaptation of Weighting Factor μ_0

In the discussion up till now, it has been assumed that the PDF and CDF of the S-R and all R-D channels are perfectly known, and hence off-line calculation of μ_0 is done. However, this might not be always feasible in practice. Therefore, we propose an alternative approach to obtain the source weighting factor, μ_0, based on real time channel measurements similar to the one adopted in [36] in the context of maintaining fairness constraints. The scheduling criteria remains the same as in (24). The source weight μ_0 is updated with each time index n according to

$$\mu_0(n+1) = \mu_0(n) + \frac{\delta}{n+1} \left(\sum_{i=1}^{M} \bar{R}_i(n+1) - \bar{R}_0(n+1) \right), \quad (37)$$

where δ controls the convergence speed of the algorithm. In our numerical results, we used $\delta = 2$. The complete algorithm is given in algorithm 2.

Algorithm 2 Real Time Weight Update

Initialize $\bar{R}_0 = 0$
Initialize $\bar{R}_i = 0 \ \ \forall \ i$
Initialize $\mu_0^{'} = \mu_{initial}$
Set μ_is to their assigned values
for n=1 to N **do**
$R_i = \log_2 \left(1 + \frac{\bar{P}|h_i|^2}{N_0} \right)$
$R_0 = \log_2 \left(1 + \frac{\bar{P}|h_{sr}|^2}{N_0} \right)$
$\mu_i^{'}(n) = (\mu_i(n) - \mu_0^{'}(n))^+ \qquad \forall \ i$
$[R^*, ind] = \max(\mu_j^{'} R_j) \ \ j = 0, 1, 2, \ldots, M$
$R_{ind} = R^*$
$R_k = 0 \ \ \forall \ k \neq ind$
$\bar{R}_j(n+1) = \bar{R}_j(n) + \frac{1}{n+1} \left(R_j - \bar{R}_j(n) \right) \ \ \forall \ j$
$\mu_0^{'}(n+1) = \mu_0^{'}(n) + \frac{\delta}{n+1} \left(\sum_{i=1}^{M} \bar{R}_i(n+1) - \bar{R}_0(n+1) \right)$
end for

D. Multiple Users Scheduling Via Superposition Coding with Successive Interference Cancellation

Up till now, single user selection per RU has been assumed. However, from (23), one can see that to achieve optimal performance, more than one user can be scheduled in a single RU. It is widely known that the optimal scheduling scheme for block-fading broadcast channels is superposition coding (SC) with successive interference cancellation (SIC) at the receivers [27], [28]. Therefore, the optimal solution is to schedule either the source (i.e. first hop) or the SC of many users (i.e. second hop) in each RU as shown in Fig. 4(b). The source and the users cannot be scheduled in the same RU because the relay operates in half-duplex mode. To accomplish the task of scheduling, the scheduler has to find the summed weighted instantaneous rates of the users and the weighted instantaneous rate of the source in the kth RU, and to schedule the one which is greater. Hence, the source is scheduled when

$$\mu_0 R_0[k] > \sum_{i=1}^{M} \mu_i R_i[k]. \quad (38)$$

Similarly, the SC of the users is selected when

$$\sum_{i=1}^{M} \mu_i R_i[k] > \mu_0 R_0[k], \quad (39)$$

where the instantaneous source rate in the kth RU can be obtained from (1). On the other hand, the procedure to calculate the instantaneous rates for the users under SC has been detailed in the literature [27]. We provide a summary of the procedure. Notice that under SC with SIC the number of users which are scheduled varies depending on the channel conditions of the users. Therefore, it is not that all users are scheduled all together, but rather a combination of some users who have good channel conditions. The order of SIC at the receivers (i.e. users) is in a decreasing order of μ, meaning that each user decodes the codewords that are sent to the users of higher μ before decoding its own codeword.

A procedure to optimize the power and rate allocation for the SC based on a greedy algorithm that involves marginal utility functions for every user was suggested in [27] for the case optimal power allocation over all channel blocks (i.e. RUs). The straightforward extension to the constant power per RU case was outlined in [35]. We summarize the latter case as follows; The marginal utility functions in the kth RU are

defined for every user as

$$u_i(z) \equiv \frac{\mu_i}{\frac{1}{|h_i[k]|^2} + z}, \qquad 0 \leq z \leq \frac{\bar{P}}{N_0}, \tag{40}$$

where z refers to the interference levels. The power and rate allocated to each user depends on the marginal utility functions according to

$$P_i[k] = N_0 \int_{\mathbb{A}_i} dz, \tag{41}$$

$$R_i[k] = \frac{1}{\ln 2} \int_{\mathbb{A}_i} \frac{1}{\frac{1}{|h_i[k]|^2} + z} dz, \tag{42}$$

where the period $\mathbb{A}_i$ is defined as

$$\mathbb{A}_i \equiv \left\{ z \in \left[0, \frac{\bar{P}}{N_0}\right] : u_i(z) > u_l(z) \ \forall l \neq i \right\}. \tag{43}$$

The user whose marginal utility function is not the maximum one over all values of $z \in \left[0, \frac{\bar{P}}{N_0}\right]$ is not scheduled in that particular RU. From (42) we obtain the achievable rate of every user, and hence we can decide whether the source should transmit or the relay should transmit using SC.

Obtaining a compact closed form expression for the long term average throughput of the source ($\bar{R}_0$) and of the users ($\bar{R}_i$) is not feasible due to the greedy algorithm to obtain the users' rates. Therefore, we obtain it in this case through numerical simulations. Similar to the single user selection case, μ_0 is obtained from the stability constraint. Hence, the optimal value of μ_0 is obtained by a numerical bisection search.

E. Proportional Fair Scheduling

We consider PF scheduling scheme for joint scheduling as we did for multi-user scheduling. The recursive algorithm is similar to Algorithm 1. However, the selection criteria changes and the source becomes considered with the users for scheduling over all RUs. For single user selection, the scheduling criteria and the update of $\mu_j^{'}$ are the same as Algorithm 2. The update equation for μ_0 is given by (37). For SC, the update equation for μ_0 is still (37). However, the instantaneous rates in each RU are calculated using the greedy procedure outlined in Section V-D.

VI. Numerical Results

We present numerical results to the different schemes discussed in this paper. We compare joint scheduling (in Section V), which offers full MHD and MUD gains, with the users-only scheduling scheme (in Section IV) and round robin scheduling scheme (in Section III). We commonly use the long term source throughput in the comparisons, which does also refer to the sum throughput of all users. We use Rayleigh block-fading channel model for the S-R and all R-D channels.

We start by considering the case of independent and identically distributed (IID) channels for all links in the first and second hops. We consider the case of maximum sum throughput scheduler, and hence we set $\mu_i = 1$ for all users. Fig. 5 shows the real-time approach to obtain μ_0 and compares

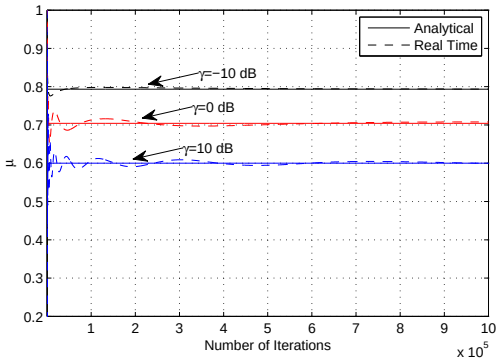

Fig. 5: Real time update of optimal μ_0 for $M = 4$.

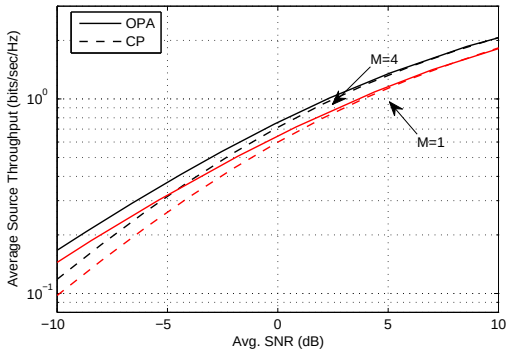

Fig. 6: Comparison between OPA and CP for joint scheduling.

it with the analytical value obtained by offline calculation from the stability constraint relating the source and user throughputs for joint scheduling with single user selection per block. We can see from Fig. 5 that the real time algorithm converges to the analytical solution. In Fig. 5, γ is defined as $\gamma = \frac{\bar{P}}{N_0}$. Moreover, the optimal value of μ_0 deceases with increasing γ. Therefore, when γ is low, the source needs to to be scheduled more often.

Fig. 6 shows a comparison between constant power (CP) and optimal power allocation (OPA) per RU for $M = 1$ and $M = 4$ with independent and identically distributed (IID) channels for joint scheduling with single user selection per block. We observe that OPA provides gains at low Avg. SNR[8] only while the gains are negligible for medium or high SNR. Therefore, applying constant power allocation is in most cases sufficient and it is also simpler to implement. Additionally, the gain for OPA decreases with increasing M. Furthermore, relaying is employed to increase the Avg. SNR of the users which have low SNR in their direct connection with the source. Hence, due to relaying, it becomes very unlikely to have users with low Avg. SNR in their connection with the relay. So, employing OPA does not grant significant gain for the system.

The long term source (sum of all users) throughput, $\bar{R}_0$, for the different schemes are plotted in Fig. 7 for four users with IID channels for the maximum sum rate scheduler. It is clear from Fig. 7 that both scenarios of joint scheduling, single user

[8]Avg. SNR=$\frac{\bar{P}\mathbb{E}[|h|^2]}{N_0}$, where $\mathbb{E}[|h|^2]$ is the average power gain of all the channels as they are taken as IID.

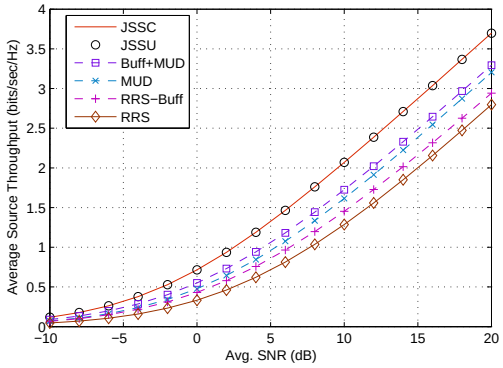

Fig. 7: Comparison of average source throughput, $\bar{R}_0$, with $M = 4$.

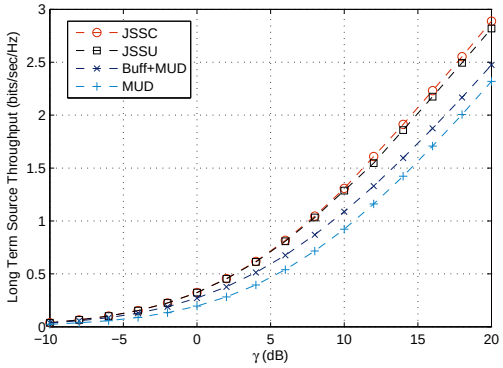

Fig. 8: Comparison of average source throughput for proportional fair scheduling, with $\bar{\gamma}_0 = 0.4$, $\bar{\gamma}_1 = 1$, and $\bar{\gamma}_2 = 0.1$.

selection (JSSU) and superposition coding (JSSC), give the best performance as we would expect due to the MHD gains. As the channels are IID and all the user weights are equal, JSSU and JSSC provide the same performance. However, the difference between them will become clear when we study PF scheduling later on. Multi-user scheduling only with buffering gives the best performance after joint scheduling as it provides both MUD and Buff gains. Furthermore, RRS with neither MHD nor MUD gains provides the worst performance as expected.

The gains of joint scheduling over users-only scheduling are also evident in the case of PF scheduling as shown in Fig. 8 which is for a two user case with different average channel values. JSSC and JSSU give comparable performance for low values of γ. However, the gain in performance due to superposition coding for JSSC becomes apparent at higher values of γ. Furthermore, both JSSC and JSSU outperform the users-only scheduling cases evidently.

The advantages of JS are also confirmed by characterizing the two-user rate region which is shown in Fig. 9. The rate region is obtained by scanning all possible user weights from $(\mu_1 = 1, \mu_2 = 0)$ to $(\mu_1 = 0, \mu_2 = 1)$. Joint scheduling enlarges the rate region due to the MHD gains. Among the JS schemes, JSSC provides a larger achievable rate region than JSSU due to scheduling multiple users in a single RU. However, the end points for both JSSC and JSSU are same as they represent the extreme cases of only one user. As expected

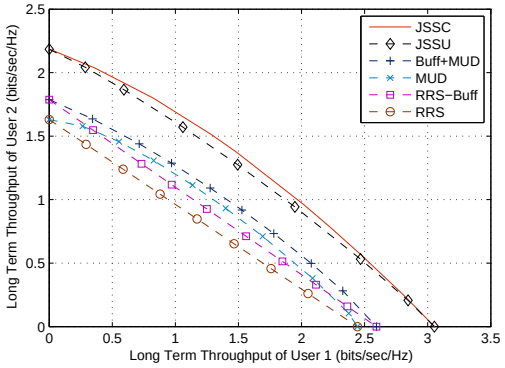

Fig. 9: Comparison of capacity regions for $\gamma = 20$ dB, with $\bar{\gamma}_0 = 0.4$, $\bar{\gamma}_1 = 1$, and $\bar{\gamma}_2 = 0.1$.

Buff+MUD gives the largest rate region after the JS schemes and RRS provides the smallest rate region. Moreover, as can be seen from Fig. 9, incorporating multi-user scheduling enhances the curvature of the rate region. However, it does not change the end points. Hence, Buff+MUD and RRS-Buff as well as MUD and RRS have the same end points respectively.

VII. Analysis of Empty Buffer Probability

The objective in this section is to get an insight into the probability of getting an empty buffer at the relay despite the fact that the scheduling scheme is optimized such that on the long-term the average rate received by the relay equals the average rate transmitted by it. The instantaneous rate transmitted or received by the relay varies with time based on the channel conditions. However, we use a simplified model, which can still give a good insight into the investigated problem. We assume that the relay transmits or receives with equal probability (equals half), and with constant rate such that the number of stored information bits in the relay's buffer changes with the same absolute value, defined as a packet, at each time index. We also assume that the initial number of packets in the relay's buffer is N. We use k for the time index. Therefore, at each k, the number of packets in the relay's buffer gets incremented (i.e. relay receives new information) or decremented (i.e. relay transmits) by one packet with equal probability.

We use K (an integer) to denote the "observation period" (we can also call it as time window), where we want to evaluate the probability that we get an empty buffer "within" this observation period (i.e. at time index $k = 1, \cdots, K$). We denote this probability as $F_E^N(K)$, where N is the initial number of packets in the buffer at $k = 0$. We use the notation K_p^N for the minimum observation period such that $F_E^N(K_p^N) \geq p$, where p is the tolerance value for the probability of the empty buffer event.

We represent the changes in the number of stored packets in the buffer within the observation period K by ω^K, which is a binary vector of K elements with values of "$+1$" or "-1". We define Ω^K as the set of all possible ω^K vectors. Therefore, the total number of elements in Ω^K is $|\Omega^K| = 2^K$. Each ω^K may occur with equal probability $p_{\omega^K} = 2^{-K}$. We associate

with each ω^K vector another vector for the accumulative changes, denoted δ_{ω^K}, which also has K elements that are obtained using $\delta_{\omega^K}(k) = \sum_{i=1}^{k} \omega^K(i)$, where $\delta_{\omega^K}(i)$ and $\omega^K(i)$ represent the ith element of δ_{ω^K} and ω^K, respectively.

The set of all ω^K vectors that result in the occurrence of the empty buffer event, given that the initial state of the buffer is N packets, is denoted Φ_N^K, where

$$\Phi_N^K = \left\{ \omega^K \ : \ \min_k \ \delta_{\omega^K}(k) \ \leq -N \ , \ \text{where } k = 1, \cdots, K \right\} \tag{44}$$

Therefore, the probability of getting an empty buffer within observation period K equals

$$F_E^N(K) = \frac{|\Phi_N^K| \ p_{\omega^K}}{|\Omega^K| \ p_{\omega^K}} = |\Phi_N^K| \ 2^{-K} \tag{45}$$

Thus, we need to compute $|\Phi_N^K|$ in order to obtain $F_E^N(K)$. This involves computing all possible combinations of ω^K that result in the empty buffer event. We can classify those into two categories. The first one is the set of ω^K vectors that have $\delta_{\omega^K}(K) \leq -N$, and the second one is the set of ω^K vectors that have $\delta_{\omega^K}(K) > -N$ and $\min_k \delta_{\omega^K}(k) \leq -N$, where $k = 1, \cdots, K-1$. We use the following two Lemmas to characterize these two categories.

Lemma 1 (Binomial Coefficients)**.** *The number of ω^K vectors that have $\delta_{\omega^K}(K) = \tilde{N}$, where $\tilde{N}$ is integer, equals zero if $|\tilde{N}| > K$ or $(K - \tilde{N})$ is odd number, and it equals $\binom{K}{(K-|\tilde{N}|)/2}$ otherwise.*

With the aid of Lemma 1, we can count the number of ω^K vectors that have $\delta_{\omega^K}(K) \leq -N$.

Lemma 2 (Equal Number in The Two Categories)**.** *For the set of ω^K vectors that have $\delta_{\omega^K}(K) < -N$, there is an equal number of ω^K vectors that have $\delta_{\omega^K}(K) > -N$ and $\min_k \delta_{\omega^K}(k) \leq -N$.*

Notice that for each ω^K that has $\delta_{\omega^K}(K) = \tilde{N}$, where $\tilde{N} < -N$, we know that we must have $\delta_{\omega^K}(k_E) = -N$ at some value $k_E < K$. Therefore, there is another ω^K which has exactly the same first k_E elements (which result in an empty buffer) and has exactly the opposite sign for the last $(K - k_E)$ elements, which results in $\delta_{\omega^K}(K) = -N - \left(\tilde{N} - (-N)\right) = -2N - \tilde{N} > -N$.

From the two Lemmas, we can obtain the closed-form expression for $F_E^N(K)$ given in (46). We can show that $F_E^N(K) = F_E^N(K-1)$ when $(K - N)$ is odd number.

After obtaining the closed-form expression to characterize $F_E^N(K)$, we can use it to obtain K_p^N. We show in Fig. 10 numerical results for K_p^N as a function of N. The figure has a log-log scale to show that the relation between K_p^N and N is approximately quadratic (linear in log-log scale with slope equals two). This means that by doubling the initial number of packets N, we get almost four times the observation period that will generate the same probability of empty buffer.

The provided analysis demonstrates that having an initial stage of source transmission only, such that the relay's buffer gets occupied by a sufficient amount of packets, is an effective way to reduce the probability of getting the buffer empty significantly. Furthermore, we believe that the empty buffer scenario does not cause a critical issue that renders the proposed joint user-and-hop scheduler impractical for implementation. Notice that even if the empty buffer scenario occurs (happens with low probability), it should not cause a major problem in practice since the system can start a new stage of source transmission only for short period of time, and then goes back to continue normal scheduling process. Therefore, the solution of this case when it happens is feasible and simple. Another remark is that, for a finite size buffer, the analysis of the probability of overloaded buffer is similar to the analysis of empty buffer.

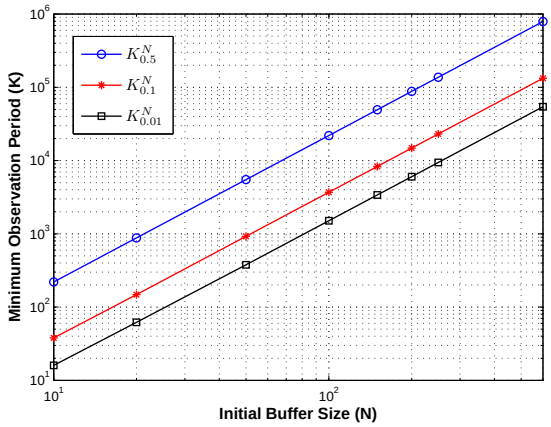

Fig. 10: K_p^N versus N for $p = 0.5$, $p = 0.1$, and $p = 0.01$.

VIII. Conclusions

We have proposed and investigated joint scheduling for dual-hop block-fading broadcast channels to maximize the weighted sum of the long term user throughputs. We have compared joint scheduling to conventional schedulers with round robin scheduling and multi-user scheduling only and we have demonstrated via numerical examples that joint scheduling provides performance improvement in terms of long term achievable rate region. We have also proposed recursive algorithms to obtain the source and user weights for proportional fair scheduling for the cases of joint scheduling and multi-user scheduling. The gains of joint user-and-hop scheduling over user-only scheduling are obtained because it exploits multi-hop diversity in addition to multi-user diversity. Due to these gains, we believe that our proposed scheduling scheme, including both cases of single user selection and superposition coding with successive interference cancellation, is favorable for practical implementation in next generation wireless systems that deploy relays such as in LTE-Advanced systems.

References

[1] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "Joint scheduling for dual-hop block-fading broadcast channels," in *Proceedings* 8^{th}

$$F_E^N(K) = \begin{cases} 0 & : \quad K < N \\ \left[\binom{K}{(K-N)/2} + 2\sum_{i=0}^{(K-N-2)/2}\binom{K}{i}\right] 2^{-K} & : \quad K \geq N \text{ and } (K-N) \text{ is even number} \\ \left[2\sum_{i=0}^{(K-N-1)/2}\binom{K}{i}\right] 2^{-K} & : \quad K > N \text{ and } (K-N) \text{ is odd number} \end{cases} \tag{46}$$

IEEE Broadband Wireless Access Workshop (BWA 2012) in conjunction with IEEE Global Communications Conference (GlobeCom 2012), Anaheim, CA, USA, Dec. 2012, pp. 140–145.

[2] E. Van der Meulen, "Transmission of information in a T-terminal discrete memoryless channel," Ph.D. dissertation, University of California, Berkeley, CA, 1968.

[3] ——, "Three-terminal communication channels," *Advances in Applied Probability*, vol. 3, pp. 120–154, 1971.

[4] T. Cover and A. Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, Sep. 1979.

[5] J. Laneman and G. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.

[6] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity-Part I: System description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[7] ——, "User cooperation diversity-Part II: Implementation aspects and performance analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, Nov. 2003.

[8] J. Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[9] R. Pabst, B. Walke, D. Schultz, P. Herhold, H. Yanikomeroglu, S. Mukherjee, H. Viswanathan, M. Lott, W. Zirwas, M. Dohler, H. Aghvami, D. Falconer, and G. Fettweis, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Communications Magazine*, vol. 42, no. 9, pp. 80–89, Sep. 2004.

[10] I. P802.16j, "IEEE standard for local and metropolitan area networks, part 16: Air interface for fixed and mobile broadband wireless access systems: Multihop relay specification," vol. P802.16j/D3, Feb. 2008.

[11] Y. Yang, H. Hu, J. Xu, and G. Mao, "Relay technologies for WiMAX and LTE-Advanced mobile systems," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 100–105, Oct. 2009.

[12] H. Ekstrom, A. Furuskar, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, and M. Wahlqvist, "Technical solutions for the 3G long-term evolution," *IEEE Communications Magazine*, vol. 44, no. 3, pp. 38–45, Mar. 2006.

[13] S. Sesia, I. Toufik, and M. Baker, *LTE-the UMTS long term evolution: from theory to practice*. John Wiley & Sons Ltd., 2011.

[14] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G evolution: HSPA and LTE for mobile broadband*. Academic Press, Elseiver, 2008.

[15] H. Ekstrom, "QoS control in the 3GPP evolved packet system," *IEEE Communications Magazine*, vol. 47, no. 2, pp. 76–83, Feb. 2009.

[16] M. Hasna and M. Alouini, "A performance study of dual-hop transmissions with fixed gain relays," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1963–1968, Nov. 2004.

[17] M. Shaqfeh and H. Alnuweiri, "Joint power and resource allocation for block-fading relay-assisted broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 10, no. 6, pp. 1904–1913, Jun. 2011.

[18] W. Mesbah and T. Davidson, "Power and resource allocation for orthogonal multiple access relay systems," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–15, article ID 476125.

[19] Y. Liang and V. Veeravalli, "Gaussian orthogonal relay channels: Optimal resource allocation and capacity," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3284–3289, Sep. 2005.

[20] Y. Liang, V. Veeravalli, and H. Poor, "Resource allocation for wireless fading relay channels: Max-min solution," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3432–3453, Oct. 2007.

[21] A. Reznik, S. Kulkarni, and S. Verdú, "Broadcast-relay channel: capacity region bounds," in *Proceedings IEEE International Symposium on Information Theory (ISIT 2005)*, Adelaide, Australia, Sep. 2005, pp. 820–824.

[22] Y. Liang and V. Veeravalli, "Cooperative relay broadcast channels," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 900–928, Mar. 2007.

[23] Y. Zou, Y. D. Yao, and B. Zheng, "Opportunistic distributed space-time coding for decode-and-forward cooperation systems," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1766–1781, Apr. 2012.

[24] L. Yang and M. Alouini, "Performance analysis of multiuser selection diversity," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 6, pp. 1848–1861, Nov. 2006.

[25] J. Boyer, D. Falconer, and H. Yanikomeroglu, "Multihop diversity in wireless relaying channels," *IEEE Transactions on Communications*, vol. 52, no. 10, pp. 1820–1830, Oct. 2004.

[26] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[27] D. N. C. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," available: www.eecs.berkeley.edu/∼dtse/broadcast2.pdf.

[28] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "On multiple users scheduling using superposition coding over Rayleigh fading channels," *IEEE Communications Letters*, to appear 2013, available: http://hdl.handle.net/10754/269853.

[29] E. L. Hahne, "Round robin scheduling for fair flow control in data communication networks," Ph.D. dissertation, Massachusetts Institue of Technology, Dec. 1986, available: http://hdl.handle.net/1721.1/14932.

[30] ——, "Round-robin scheduling for max-min fairness in data networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 1024–1039, Jul. 1991.

[31] M. Shaqfeh and N. Goertz, "Performance analysis of scheduling policies for delay-tolerant applications in centralized wireless networks," in *Proceedings IEEE International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2008)*, Edinburgh, UK, Jun. 2008, pp. 309–316.

[32] M. Shaqfeh, N. Goertz, and S. McLaughlin, "Organizing multiuser operation in centralized wireless networks," in *Wireless World Research Forum Meeting 20*, Ottawa, Canada, Apr. 2008, pp. 1–6.

[33] B. Hogstad, M. Patzold, N. Youssef, and V. Kontorovitch, "Exact closed-form expressions for the distribution, the level-crossing rate, and the average duration of fades of the capacity of OSTBC-MIMO channels," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 1011–1016, Feb. 2009.

[34] M. Shaqfeh and N. Goertz, "Comments on the boundary of the capacity region of multiaccess fading channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3407–3408, Jul. 2009.

[35] M. Shaqfeh, N. Goertz, and J. Thompson, "Ergodic capacity of block-fading Gaussian broadcast and multi-access channels for single-user-selection and constant-power," in *Proceedings* 17^{th} *European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, UK, Aug. 2009, pp. 784–788.

[36] M. Shaqfeh and N. Goertz, "Channel-aware scheduling with resource-sharing constraints in wireless networks," in *Proceedings IEEE International Conference on Communications (ICC 2008)*, Beijing, China, Jun. 2008, pp. 4149–4153.

Ammar Zafar (S'12) received his B.E. and M.Sc. degrees in electrical engineering from National University of Sciences and Technology (NUST), Pakistan, in 2007 and 2009, respectively. He joined the Department of Electrical Engineering at King Abdullah University of Science and Technology (KAUST) in August 2010, where he is currently a Ph.D student. His research interests include wireless communication theory, cognitive radio networks, cooperative networks and multiuser scheduling.

Mohammad Obaidah Shaqfeh (S'07-M'09) received the B.Sc. degree in electrical engineering (communications stream) from United Arab Emirates University in 2003 and the M.Sc. degree in communications technology from Ulm University, Germany in 2005. He received the Ph.D. degree from The University of Edinburgh, Edinburgh, Scotland, U.K., in 2009. In January 2009, Dr. Shaqfeh joined the Department of Electrical and Computer Engineering at Texas A&M University at Qatar, where he is currently working as an Associate Research Scientist. His research interests include wireless communications systems and information theory. In particular, he is interested in layered transmission schemes, relay-assisted communication, optimal resource allocation, and multiuser scheduling.

Mohamed-Slim Alouini (S'94, M'98, SM'03, F'09) was born in Tunis, Tunisia. He received the Ph.D. degree in Electrical Engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He served as a faculty member in the University of Minnesota, Minneapolis, MN, USA, then in the Texas A&M University at Qatar, Education City, Doha, Qatar before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia as a Professor of Electrical Engineering in 2009. His current research interests include the modeling, design, and performance analysis of wireless communication systems.

Hussein M. Alnuweiri (S'81–M'83) received the Master's degree from King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia in 1984, and the Ph.D. degree in electrical and computer engineering from the University of Southern California, Los Angeles in 1989. He is currently a Professor and Program Chair at the Department of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar. From 1991 to 2007, he was a Professor with the Department of Electrical and Computer Engineering, University of British Columbia. From 1996 to 1998, he also represented the University of British Columbia, Vancouver, BC, Canada, at the ATM Forum. From 2000 to 2006, he served as a Canadian delegate to the ISO/IEC JTC1/SC29 Standards Committee (MPEG-4 Multimedia Delivery), where he worked within the MPEG-4 standardization JTC1-SC29WG11 and the Ad-Hoc group involved in the development of the reference software IM1 AHG.

Dr. Alnuweiri has a long record of industrial collaborations with several major companies worldwide. He is also an inventor, and holds three U.S. patents, and one International patent. He has authored or co-authored over 150 refereed journal and conference papers in various areas of computer and communications research. In particular, his research interests include mobile Internet technologies, multimedia communications, wireless protocols, routing and information dissemination algorithms for opportunistic networking, and quality-of-service provisioning and resource allocation in wireless networks.