Resource Allocation for Two Source-Destination Pairs Sharing a Single Relay with a Buffer

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

Abstract—In this paper, we obtain the optimal resource allocation scheme in order to maximize the achievable rate region in a dual-hop system that consists of two independent source-destination pairs sharing a single half-duplex relay. The relay decodes the received information and possesses buffers to enable storing the information temporarily before forwarding it to the respective destination. We consider both non-orthogonal transmission with successive interference cancellation at the receivers and orthogonal transmission. Also, we consider Gaussian block-fading channels and we assume that the channel state information is known and that no delay constraints are required. We show that, with the aid of buffering at the relay, joint user-and-hop scheduling is optimal and can enhance the achievable rate significantly. This is due to the joint exploitation of multiuser diversity and multihop diversity in the system. We provide closed-form expressions to characterize the average achievable rates in a generic form as functions of the statistical model of the channels. Furthermore, we consider sub-optimal schemes that exploit the diversity in the system partially and we provide numerical results to compare the different schemes and demonstrate the gains of the optimal one.

Index Terms—Optimal resource allocation, achievable rate region, block-fading channels, shared relay, relay with a buffer, joint user-and-hop scheduling.

I. INTRODUCTION

We aim in this introductory section to give a clear picture of the objectives and subject of this paper by emphasizing three keywords; "resource allocation", "shared relay" and "relay with a buffer".

A. Theme: Optimal Resource Allocation

It is well-known that next-generation wireless systems are prospected to provide large gains in the supported data rate targets and quality-of-service constraints. The major bottleneck is that the air-link (i.e. bandwidth) resources over the wireless medium are limited, precious and shared by all service providers and respective users. Therefore, enhancing the

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).

resource allocation efficiency is not only a privilege, it is also a must. Furthermore, the emerging communication systems, such as Long-Term-Evolution (LTE) [1]–[3], are aided by adaptive transmission schemes (i.e. modulation and coding) and dynamic resource allocation and multiuser scheduling methods [4]. Therefore, optimal resource allocation is both essential and applied in practice.

The formulation of the resource allocation optimization problem takes variant forms in the literature depending on the used objective functions and constraints. In this work, we aim to maximize the achievable rate region over Gaussian block-fading channels and we use the best transmission schemes that maximize the achievable rates. These require non-orthogonal transmission with successive interference cancellation at the receivers. We do additionally consider and compare with orthogonal transmission. Furthermore, we assume best-effort data traffic without delay constraints such that the primal objective is to maximize a weighted sum (over the users) of the average (over the channel blocks) achievable rates. We assume adaptive rate transmission based on the channel conditions, which is known at the receiver and at the transmitter via feedback. Notice that some criteria like minimum rate requirements, maximum delay tolerance or outage probability are not within the scope of our work. Our approach to the resource allocation problem is similar to some works in the literature such as [5] for the point-to-point channel, [6] for the multi-access (many-to-one) channels, [7], [8] for the broadcast (one-to-many) channels, [9] for a reduced-feedback scheme over broadcast channels, [10] for the relay channels, [11] for the relay-assisted broadcast channel, and our recent work [12] on the dual-hop broadcast channel with a buffer at the relay.

We would like to emphasize that although the approach and problem formulation is similar in the aforementioned papers, each channel model produces its unique optimization problem with different constraints and solution steps. Therefore, in general, the solution of a new channel model is not a simple extension of other existing solutions. That's why we have a rich literature on optimal resource allocation. In this paper, we consider a new channel setup which, to the best of our knowledge, has not been solved in the literature yet. Next, we describe our targeted channel model and its relevance in practice.

B. Subject: Shared Relay System

In the last few years, there has been a growing research interest in exploiting the broadcast nature of the wireless medium and the potential gains of the interaction (cooperation) between neighboring nodes [13], [14]. Furthermore, relaying technologies have been designated to be major new enabling technologies in the wireless standards. For instance, fixed access-point relays are to be deployed in next generation wireless systems, such as LTE-Advanced [15], [16]. The current trend in the communication industry is to focus on innovative technical solutions at the network architecture level because the technical solutions which are based on the link level (i.e. point-to-point communication) are well matured and advanced. Relaying technologies are demonstrated to enhance the coverage and transmission capacity of wireless networks. Notice that the relays do not have wired connection to the backhaul network in LTE-Advanced systems. Furthermore, users cooperation (e.g. a mobile terminal acts as a relay to forward the information sent by other mobile terminals) is not considered in LTE-Advanced systems although it is a sort of cooperative communication. Therefore, our interest in this work is mainly on relay-assisted communication.

In the literature, the relay channel was first proposed in the 1960s [17]–[19], and the capacity of the relay channel from information-theoretic perspective was characterized in [20]. More extensive research efforts in the topic of relaying / cooperative communication have been revived in the last decade. Some examples of important contributions in this topic include [21]–[41].

In this work, we consider a half-duplex shared relay channel, which corresponds to the case when a single relay serves two adjacent wireless cells (i.e. base stations). Thus, our system model consists of two sources which communicate with two destinations (i.e. users) independently using dual-hop connections through a single shared relay. We believe that this is an important system model which can be relevant in the practical deployment of relays in cellular systems. The shared relay channel is an interesting model that combines a multiple-access channel, a broadcast channel, and a relay channel altogether. Therefore, good background knowledge about all of them is needed in order to treat the shared relay model. Additionally, we consider an important factor to get considerable gains in the achievable rates, which is by enabling information buffering (i.e. temporary storage) at the relay node. This is an important feature that has not been studied in the literature comprehensively yet. Therefore, we would like to emphasize it more.

C. Important Feature: Relay with a Buffer

The use of a buffer in a relay has been demonstrated to enhance the throughput over a three-node (source, destination and relay) channel in [42], [43]. Moreover, the buffers improve the performance of relay selection as discussed in [44]–[46]. More relevant to our specific problem in this paper, we have shown in [12], [47] that buffering enables exploiting the diversity in the two hops through a relay to obtain significant capacity (i.e. achievable rate) gains. The multihop diversity [48] is due to the independent fading conditions over the two hops. Therefore, it becomes more likely that at least one of them will have very good channel conditions. Thus, the capacity gains can be obtained by hop-scheduling based on the channel conditions. It is obvious that without buffering, the relay has to forward the received information from a source to the respective destination directly regardless of the channel conditions of the second hop. Thus, we can think of hop-scheduling as a decoupling of the first hop and second hop so that if one of them has bad channel condition, the transmission rate over the other one does not get degraded consequently as well. More interestingly, we have discussed in [12] how to integrate user-and-hop scheduling in order to exploit both multiuser diversity (MUD) [49] and multihop diversity (MHD) jointly. The channel model in [12] consists of a dual-hop broadcast channel with a single source serving multiple users through a single relay. The extension of the joint user-and-hop scheduling to the shared relay model is intuitively more desirable especially that we have more diversity at the first hop because there are two sources in this case.

The technical challenge in joint user-and-hop scheduling is the need to maximize the throughput while taking into consideration the fundamental law of the conservation of flow (rate) at the queue of a buffer, which is well known in the queuing theory. Therefore, the optimal scheduling scheme should make the average rate received by the relay matches the average rate transmitted by it. We would like to mention that we assume that the sources have infinite backlog such that they can always transmit as much information as their channels' capacity and that the buffers of the relay have infinite size.

D. Summary of Our Contributions

In this paper, we show how to extend the joint user-and-hop scheduling scheme into a shared relay with a buffer. In particular, we have the following contributions in this work:

- We characterize the optimal resource allocation strategies in order to maximize the achievable rate region for our system over Gaussian block-fading channels. We illustrate also how to obtain the optimal solution using an off-line approach that is based on the knowledge of the channel statistics or using a real-time approach that is based on actual channel measurements.
- We derive closed-form formulas to characterize the achievable rates in our system as a function of the fading channels statistics (mainly the probability density function). This is a generic solution that is applicable to any specific fading channel model.
- We investigate sub-optimal schemes by neglecting the buffer at the relay, and by adding orthogonality constraint to the problem. We compare the sub-optimal schemes with the optimal one numerically in order to characterize the capacity gains due to buffering and non-orthogonal transmission. Additionally, we compare with other sub-optimal schemes that exploit the diversity gains partially.

- We demonstrate in the numerical results that the gains due to buffering at the relay are significant.

The rest of the paper is organized as follows. We elaborate more on the system model in Section II and we formulate the main optimization problem. We provide the optimal solution in Section III. Next, we discuss the solution when buffering is not used in Section IV. After that, we characterize sub-optimal schemes in Section V. Next, we provide various numerical results in Section VI. Finally, we summarize the main conclusions of the paper in Section VII.

II. System Model and Problem Formulation
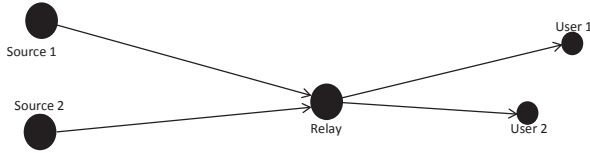
A. System Model and Assumptions

Fig. 1: System Model: Dual-hop system with two source-destination pairs.

We consider the system shown in Fig. 1, where two sources (S_1 and S_2) send information to two users (destinations D_1 and D_2) through a single shared relay (R). In this system, S_1 sends information to D_1 only and S_2 sends information to D_2 only. There are no direct links between the two sources and the two destinations. Hence, all sent information must go through the relay. The relay decodes[1] the received information from the sources and it has two buffers (one for each source) to enable storing the information temporarily before forwarding it to the respective destination. We assume that the amount of information that can be stored in the buffers is unlimited. When the relay is scheduled to transmit, information is extracted from each buffer and then mapped onto codewords with rates, $R_{D1}[k]$ and $R_{D2}[k]$, as explained later on. Furthermore, we assume that each node is equipped with a single antenna. Moreover, the relay cannot receive and transmit simultaneously in a given frequency band, i.e. it works in a half-duplex mode.

We assume a Gaussian block-fading (i.e. quasi-static) channel model in which the time-frequency grid is divided into blocks [51]–[53]. Each channel block, called a resource unit (RU), can be composed of multiple time slots and multiple frequency bands. However, the RUs are orthogonal to each other. The complex channel gains of all links, S-R and R-D, stay constant over a single RU and they change independently from one RU to another. Furthermore, all links are independent of one another. We assume, without loss of generality, that the RUs have the same duration and bandwidth.

Additionally, we assume that the relay uses a constant power spectral density (in Jouls/sec/Hz) whenever it transmits. Furthermore, the sources use a constant sum (of the two sources) power spectral density whenever one or both of them transmit. The channel gains of the four links in the system are known for each RU and they are incorporated in the resource allocation for every RU. We can assume that the scheduling decisions are taken at the relay since it has connections with all sources and destinations. Thus, all channel state information (CSI) can be collected via single-hop connections instead of dual-hop connections. The relay can, for example, transmit a pilot signal and then all other nodes measure their channels' strengths and send the CSI via feedback to the relay. With the assumption of channel reciprocity, the relay can in this case have the CSI of all channels and then it can broadcast the resource allocation decision to all nodes so that they all become aware of it and adapt accordingly.

B. Optimization Problem Formulation

The primal objective function for resource allocation is to maximize the *weighted sum of the average achievable rates* (in bits/sec/Hz) of the two users (D_1 and D_2) over a large number of channel blocks (i.e. RUs). There are no delay constraints involved in this problem and hence the resource allocation is based merely on the channel state information and the predefined (i.e. given) weights[2] for the two users. We can formulate the optimization problem as

$$\max\Big(\mu_1 \min(\bar{R}_{S1}, \bar{R}_{D1}) + \mu_2 \min(\bar{R}_{S2}, \bar{R}_{D2})\Big), \quad (1)$$

where μ_1 and μ_2 are given weights. Also, $\bar{R}_{S1}$ and $\bar{R}_{S2}$ are the average rates transmitted by S_1 and S_2, respectively, and decoded reliably at the relay. Similarly, $\bar{R}_{D1}$ and $\bar{R}_{D2}$ are the average rates transmitted by the relay and decoded reliably at D_1 and D_2, respectively. The averaging is over all RUs. We give index k for the resource units. Therefore, we can write

$$\bar{R}_{S1} = \frac{1}{K}\sum_{k=1}^{K} R_{S1}[k], \quad (2)$$

where $R_{S1}[k]$ is the rate transmitted by S_1 in the kth RU, and K is the total number of RUs, which is assumed to be very large. The same notation can be applied to $\bar{R}_{S2}$, $\bar{R}_{D1}$ and $\bar{R}_{D2}$. The optimization variables in problem (1) are the resource allocation decisions over the RUs. We define these variables specifically for both the optimal non-orthogonal transmission case and the sub-optimal (but preferable in practice) orthogonal transmission case in Section II-C. We use $\xi[k]$ to denote the vector of optimization variables for the kth RU.

The minimum of $\bar{R}_{S1}$ and $\bar{R}_{D1}$ is taken in (1) since the information sent by S_1 to D_1 goes over dual hops and hence the rate is bounded by the worst hop. This is also the case for

---

[1]Notice that over dual-hop channel, decode-and-forward is the optimal relaying strategy and it achieves the capacity. However, for a relay channel with direct link from the source the destination, compress-and-forward can be better than decode-and-forward in some cases and the capacity is not known in general [50]. This case is not relevant to our problem since we assume no direct links between the sources and the destinations.

[2]As discussed in [54], [55], the specific selection of the users' weights control the specific "pareto-optimal" operating point on the boundary of the achievable rate region of the system. Therefore, the weights can be adjusted to maintain fairness among the users or achieve specific average rate requirements.

$\bar{R}_{S2}$ and $\bar{R}_{D2}$. It is intuitive that the optimal solution of (1) must involve transmitting at channel capacity over each RU, which is characterized in Section II-C. Therefore, we assume that the sources and the relay uses capacity achieving channel codes. Furthermore, the transmission rate of the relay to the destination is also bounded by the amount of the information bits that is stored in the queues of its two buffers. Thus, the relay may not be able to transmit at channel capacity in some RUs consequently.

Moreover, we must distribute the resources such that $\bar{R}_{S1} = \bar{R}_{D1}$ and $\bar{R}_{S2} = \bar{R}_{D2}$. Otherwise, we can predict that some of the resources allocated to one hop are wasted if it has useless extra rate than the other hop. Therefore, we can write (1) in the equivalent form

$$\max_{\xi[k], \forall k} \quad \mu_1 \bar{R}_{D1} + \mu_2 \bar{R}_{D2} \tag{3a}$$

$$\text{subject to} \quad \bar{R}_{D1} \le \bar{R}_{S1} \quad \text{and} \quad \bar{R}_{D2} \le \bar{R}_{S2}, \tag{3b}$$

where $\forall k$ stands for $k \in \{1, 2, \ldots, K\}$. We refer to (3b) as the flow conservation constraint because it guarantees that on the long-term the average rates input to the relay's two buffers match the average rates output from them[3]. However, it does not put any constraint on how long the information can stay in the relay's buffers before forwarding it. This feature enables exploiting the diversity in the four links in the system to maximize the average rates. This is feasible since we have no delay constraints in our problem and the sizes of the two buffers of the relay are unbounded.

In the sub-optimal case when buffering is not supported, the relay has to forward the received information immediately to the respective user in the same RU. Therefore, in this case, the constraint (3b) should be replaced by K constraints,

$$R_{D1}[k] = R_{S1}[k] \quad \text{and} \quad R_{D2}[k] = R_{S2}[k] \qquad \forall k. \tag{4}$$

C. Transmission Schemes and Optimization Variables

Since the relay is half-duplex, each RU is divided orthogonally in the time domain into two sub-blocks, where the relay receives information from the sources in the first one, and transmits to the destinations in the second one. The ratios of the time durations of first and second sub-blocks in the kth RU to the total time of the RU are denoted $\tau_S[k]$ and $\tau_D[k]$, respectively. Therefore, we can write

$$R_{S1}[k] = \tau_S[k] R_{S1}^{ac}[k] \tag{5a}$$

$$R_{D1}[k] = \tau_D[k] \min\left(Q_1[k], R_{D1}^{ac}[k]\right), \tag{5b}$$

where $R_{S1}^{ac}[k]$ is the achievable rate (capacity in bits/sec/Hz) of the channel between S_1 and the relay in the kth RU. Similarly, $R_{D1}^{ac}[k]$ is the achievable rate of the channel between the relay and D_1 in the kth RU. Also, $Q_1[k]$ is the normalized amount of information, which is stored in the queue of the buffer that

is used for the first S-D pair, at the start of the kth RU. We define $R_{S2}[k]$, $R_{S2}^{ac}[k]$, $R_{D2}[k]$, $R_{D2}^{ac}[k]$ and $Q_2[k]$ similar to (5). Notice that the relay cannot transmit more information than what is actually stored in its buffers. That's why we use (5b).

The optimal transmission schemes for the multiple-access channel from the sources to the relay is non-orthogonal transmission by the two sources. The relay should decode the codeword of the source that has the stronger channel first treating the signal of the other source as noise. Then, it decodes the codeword of the "weaker channel source" after removing the first one from the received signal[4]. For the broadcast channel from the relay to the destinations, the relay uses superposition coding for two codewords of the destinations. The destination that has the stronger channel decodes the codeword of the other destination first before it can obtain its own codeword [56]. The optimization variables for both cases are the ratios of the total power that are allocated to each codeword. We use the notation $\rho_{S1}[k]$ and $\rho_{S2}[k]$ for the ratios of the power allocated to S_1 and S_2, respectively, to the total sum power constraint of the multiple-access channel. Similarly, we use the notation $\rho_{D1}[k]$ and $\rho_{D2}[k]$ for the ratios of the power allocated to the codewords sent to D_1 and D_2, respectively, to the total power of the relay. With these notations, we can characterize $R_{S1}^{ac}[k]$ and $R_{D1}^{ac}[k]$ using

$$R_{S1}^{ac}[k] = \begin{cases} \log\left(1 + \rho_{S1}[k]\gamma_{S1}[k]\right) & : \gamma_{S1}[k] < \gamma_{S2}[k] \\ \log\left(1 + \frac{\rho_{S1}[k]\gamma_{S1}[k]}{1 + \rho_{S2}[k]\gamma_{S2}[k]}\right) & : \gamma_{S1}[k] > \gamma_{S2}[k] \end{cases} \tag{6}$$

$$R_{D1}^{ac}[k] = \begin{cases} \log\left(1 + \frac{\rho_{D1}[k]\gamma_{D1}[k]}{1 + \rho_{D2}[k]\gamma_{D1}[k]}\right) & : \gamma_{D1}[k] < \gamma_{D2}[k] \\ \log\left(1 + \rho_{D1}[k]\gamma_{D1}[k]\right) & : \gamma_{D1}[k] > \gamma_{D2}[k], \end{cases} \tag{7}$$

where γ refers to the signal-to-noise-ratio (SNR) of the respective channel. Similarly, we can characterize $R_{S2}^{ac}[k]$ and $R_{D2}^{ac}[k]$ using (6) and (7), respectively, after swapping index 1 and index 2 for the SNRs and power ratios.

The optimization variables for each RU are $\tau_S[k]$, $\tau_D[k]$, $\rho_{S1}[k]$, $\rho_{S2}[k]$, $\rho_{D1}[k]$ and $\rho_{D2}[k]$. Therefore, $\xi[k]$ in the optimization problem (3) is defined as

$$\xi[k] = \left[\tau_S[k], \tau_D[k], \rho_{S1}[k], \rho_{S2}[k], \rho_{D1}[k], \rho_{D2}[k]\right]. \tag{8}$$

The space of the vector $\xi[k]$ is defined by

$$\tau_S[k] + \tau_D[k] = 1, \quad \tau_S[k] > 0, \quad \tau_D[k] > 0, \tag{9a}$$

$$\rho_{S1}[k] + \rho_{S2}[k] = 1, \quad \rho_{S1}[k] \ge 0, \quad \rho_{S2}[k] \ge 0, \tag{9b}$$

$$\rho_{D1}[k] + \rho_{D2}[k] = 1, \quad \rho_{D1}[k] \ge 0, \quad \rho_{D2}[k] \ge 0. \tag{9c}$$

When orthogonal transmission is used, we have

$$R_{S1}^{ac}[k] = \omega_{S1}[k] \log\left(1 + \gamma_{S1}[k]\right), \tag{10}$$

---

[3]From the fundamental law of conservation of flow at the queue of a buffer, the long term average input rate to the queue of the buffer should be greater than or equal the long term average output rate. However, the optimal solution of (3) is achieved when the constraint is satisfied at equality.

[4]The capacity region of the multiple-access channel with sum power constraint is the same as of the dual broadcast channel with the opposite optimal decoding order. There is no need for time sharing in this case as there is for individual power constraints.

where $\omega_{S1}[k]$ is the (time / bandwidth) ratio allocated to S$_1$ to the total (time / bandwidth) of the multiple-access channel. Similar formulas can be used to characterize $R_{S2}^{ac}[k]$, $R_{D1}^{ac}[k]$ and $R_{D2}^{ac}[k]$. Furthermore, $\xi[k]$ in this case is defined as

$$\xi[k] = \left[\tau_S[k], \tau_D[k], \omega_{S1}[k], \omega_{S2}[k], \omega_{D1}[k], \omega_{D2}[k]\right]. \quad (11)$$

The space of the vector $\xi[k]$ in this case is defined by

$$\omega_{S1}[k] + \omega_{S2}[k] = 1, \quad \omega_{S1}[k] \geq 0, \quad \omega_{S2}[k] \geq 0, \quad (12a)$$
$$\omega_{D1}[k] + \omega_{D2}[k] = 1, \quad \omega_{D1}[k] \geq 0, \quad \omega_{D2}[k] \geq 0, \quad (12b)$$

in addition to (9a).

D. Buffers at the Edge of Non-Absorption

One technical challenge in the solution of the optimization problem and the analytical characterization of the long-term average achievable rates is that the rates on the second hop of the relay must involve the minimum of the actual amount of data stored in the queues of the two buffers and the actual channel capacities of the two relay-destination links, as shown in (5b). However, it was shown in [43, Theorem 2] that when the optimal resource allocation scheme is applied (i.e. maximizing the throughput) in a dual-hop channel with a buffer-aided relay, the buffer of the relay will be at the edge of non-absorption, meaning that the average arrival and departure rates are equal. Furthermore, it was proved therein that the average rate on the second hop would be be the same as its channel average capacity, and the case of having less amount of information stored in the queue than the channel capacity would appear very rarely and its effect on the long-term average achievable rate would be ignored when averaging over infinite number of RUs. This is valid in our problem as well and it applies to both buffers since both will be on the edge of non-absorption and at optimality their constraints in (3b) will be satisfied at equality. Consequently, using a similar proof like in [43, Theorem 2], we can show that at optimality

$$\bar{R}_{D1} = \frac{1}{K}\sum_{k=1}^{K}\tau_D[k]\min\left(Q_1[k], R_{D1}^{ac}[k]\right) = \\ \frac{1}{K}\sum_{k=1}^{K}\tau_D[k]R_{D1}^{ac}[k], \quad (13)$$

and the same applies to $\bar{R}_{D2}$.

Therefore, we obtain the optimal resource allocation in Section III based merely on channel conditions and we ignore the two queues states. Nevertheless, it should be clear that if it occasionally happens that on some RUs the relay can not send at the allocated rate for one or both destinations, it should adjust the transmission rate according to the available amount of information in its buffers.

As a final remark, the fact that the actual queue state has negligible effect on the optimal resource allocation scheme was also discussed from a different perspective in [47, Section VII], where the probability of getting an empty buffer was analyzed giving the initial state of the queue. It was demonstrated that this probability can be made arbitrary small given that the queue has sufficient amount of data initially. Therefore, we could have an apriori filling stage (i.e. transmissions from the sources to the relay only) before applying the optimal resource allocation scheme. This apriori filling stage would be relatively very short with respect to the total number of RUs.

III. OPTIMAL SOLUTION FOR BUFFER-AIDED RELAY

A. Lagrangian Dual Problem Formulation

To solve the optimization problem in (3), we use the Lagrangian dual problem [57], which yields

$$\min_{\lambda_{S1},\lambda_{S2}} \quad \Phi(\lambda_{S1}, \lambda_{S2}), \quad (14)$$

where $\lambda_{S1} \geq 0$ and $\lambda_{S2} \geq 0$ are the Lagrange multipliers corresponding to the constraints in (3b), and $\Phi(\lambda_{S1}, \lambda_{S2})$ is given by (15).

We can write (15) in terms of the rates in each RU, which yields (16). We observe that problem (16) can be divided into K independent optimization problems. This is because the instantaneous rates in the kth RU are only dependent on their respective optimization variables in the kth RU (i.e. $\xi[k]$) and are independent of the optimization variables in all other RUs. Therefore, (16) is equivalent to

$$\Phi(\lambda_{S1}, \lambda_{S2}) = \sum_{k=1}^{K}\Phi(\lambda_{S1}, \lambda_{S2})[k], \quad (20)$$

where $\Phi(\lambda_{S1}, \lambda_{S2})[k]$ is given by (17), $\lambda_{D1} = \mu_1 - \lambda_{S1}$ and $\lambda_{D2} = \mu_2 - \lambda_{S2}$.

As we will show in the sequel, the solution for (17) is always unique for all values of λ_{S1} and λ_{S2}. Consequently, the solution of (15) is also unique. Therefore, we have strong duality in our problem and the optimal solution of (14) is the optimal solution of the primal problem (3) as well. Furthermore, the solution of (14) is at λ_{S1} and λ_{S2} that satisfy the flow conservation constraints in (3b). Notice that we could have strong duality even in nonconvex optimization problems. This is known in the literature, e.g. [58]–[60].

B. Optimal Resource Allocation

Noting that the time sharing variable $\tau_S[k]$ does not affect the achievable rates, we can write (17) in the form given in (18) where $\tau_D[k] = 1 - \tau_S[k]$, and

$$\Phi_S(\lambda_{S1}, \lambda_{S2})[k] = \max_{\rho_{S1},\rho_{S2}} \quad \lambda_{S1}R_{S1}^{ac}[k] + \lambda_{S2}R_{S2}^{ac}[k], \quad (21a)$$

$$\Phi_D(\lambda_{S1}, \lambda_{S2})[k] = \max_{\rho_{D1},\rho_{D2}} \quad \lambda_{D1}R_{D1}^{ac}[k] + \lambda_{D2}R_{D2}^{ac}[k]. \quad (21b)$$

Notice that we should replace ρ by ω in (21) for the orthogonal transmission case. Furthermore, the optimal solution of $\tau_S[k]$ in (18) is

$$\Phi(\lambda_{S1}, \lambda_{S2}) = \max_{\xi[k], \forall k} \quad \mu_1 \bar{R}_{D1} + \mu_2 \bar{R}_{D2} - \lambda_{S1}(\bar{R}_{D1} - \bar{R}_{S1}) - \lambda_{S2}(\bar{R}_{D2} - \bar{R}_{S2}) \tag{15}$$

$$\Phi(\lambda_{S1}, \lambda_{S2}) = \max_{\xi[k], \forall k} \quad (\mu_1 - \lambda_{S1}) \sum_{k=1}^{K} R_{D1}[k] + (\mu_2 - \lambda_{S2}) \sum_{k=1}^{K} R_{D2}[k] + \lambda_{S1} \sum_{k=1}^{K} R_{S1}[k] + \lambda_{S2} \sum_{k=1}^{K} R_{S2}[k] \tag{16}$$

$$\Phi(\lambda_{S1}, \lambda_{S2})[k] = \max_{\xi[k]} \quad \lambda_{D1} R_{D1}[k] + \lambda_{D2} R_{D2}[k] + \lambda_{S1} R_{S1}[k] + \lambda_{S2} R_{S2}[k] \tag{17}$$

$$\Phi(\lambda_{S1}, \lambda_{S2})[k] = \max_{\tau_S[k]} \quad \tau_S[k] \Phi_S(\lambda_{S1}, \lambda_{S2})[k] + \tau_D[k] \Phi_D(\lambda_{S1}, \lambda_{S2})[k] \tag{18}$$

$$\omega_{S1}[k] = \begin{cases} 1 & \text{if} \quad \lambda_{S1} \log(1 + \gamma_{S1}[k]) > \lambda_{S2} \log(1 + \gamma_{S2}[k]) \\ 0 & \text{if} \quad \lambda_{S1} \log(1 + \gamma_{S1}[k]) < \lambda_{S2} \log(1 + \gamma_{S2}[k]) \end{cases} \tag{19a}$$

$$\omega_{D1}[k] = \begin{cases} 1 & \text{if} \quad \lambda_{D1} \log(1 + \gamma_{D1}[k]) > \lambda_{D2} \log(1 + \gamma_{D2}[k]) \\ 0 & \text{if} \quad \lambda_{D1} \log(1 + \gamma_{D1}[k]) < \lambda_{D2} \log(1 + \gamma_{D2}[k]) \end{cases} \tag{19b}$$

$$\tau_S[k] = \begin{cases} 1 & \text{if} \quad \Phi_S[k] > \Phi_D[k] \\ 0 & \text{if} \quad \Phi_S[k] < \Phi_D[k]. \end{cases} \tag{22}$$

This means that in a single RU, either the sources transmit or the relay transmits and not both of them, depending on the channel conditions of the four links of the system. That's why we like to refer to this resource allocation problem as a "scheduling" problem. Notice that the probability that $\Phi_S[k] = \Phi_D[k]$ equals zeros when the probability density functions of the channel gains are continuous functions. This is a valid assumption for our problem.

The next step is to solve (21). We start first by the **orthogonal transmission** case because it is simpler. The optimal $\omega_{S1}[k]$ and $\omega_{D1}[k]$ in this case are given by (19).

Then, it is straightforward to obtain $\omega_{S2}[k] = 1 - \omega_{S1}[k]$ and $\omega_{D2}[k] = 1 - \omega_{D1}[k]$. From (22) and (19) we find that at each RU, a single link is scheduled among the four links. So, either only a single source transmits or the relay transmits to only a single user. Therefore, we call this scheme "joint user-and-hop scheduling". Since the scheduling criterion is based mainly on the channel conditions of the two sources and two users in every RU, the system can exploit both MUD and MHD jointly to enhance the achievable rates. In a compact form, we can present the scheduling criterion as choosing $\eta[k]$, where

$$\eta[k] = \arg \max_x \lambda_x \log(1 + \gamma_x[k]), \tag{23}$$

where $x \in \{S_1, S_2, D_1, D_2\}$. Fig. 2 shows an example of the dynamic scheduling of the sources and users over the RUs. Notice that when we say that a user is scheduled, this means that the relay transmits information destined to this particular user.

For the **non-orthogonal transmission** case[5], the optimal ρ_{S1} and ρ_{D1} can be obtained by making the gradient equals zero in (21) from the KKT conditions. So, we get the following

---

[5]It is well-known that with non-orthogonal transmission, we can achieve higher capacity gains by allowing multiple (sources / users) scheduling per RU as demonstrated numerically in [61] for a broadcast channel case.
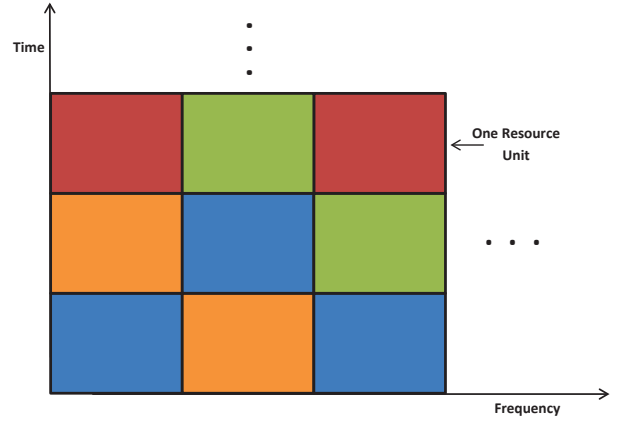

Fig. 2: Example of the RUs allocation for joint user-and-hop scheduling with single user selection per RU. Blue represents S_1, yellow represents S_2, red represents D_1 and green represents D_2.

conditions for the optimal solution:

$$\frac{\partial \Phi_S[k]}{\partial \rho_{S1}[k]} = \frac{\partial \Phi_S[k]}{\partial \rho_{S2}[k]}, \quad \rho_{S1}[k] + \rho_{S2}[k] = 1, \tag{24a}$$

$$\frac{\partial \Phi_D[k]}{\partial \rho_{D1}[k]} = \frac{\partial \Phi_D[k]}{\partial \rho_{D2}[k]}, \quad \rho_{D1}[k] + \rho_{D2}[k] = 1. \tag{24b}$$

From (24a), and with reference to (6), we get

$$\lambda_{S1} \left(\frac{\partial R_{S1}^{ac}[k]}{\partial \rho_{S1}[k]} - \frac{\partial R_{S1}^{ac}[k]}{\partial \rho_{S2}[k]} \right) = \lambda_{S2} \left(\frac{\partial R_{S2}^{ac}[k]}{\partial \rho_{S2}[k]} - \frac{\partial R_{S2}^{ac}[k]}{\partial \rho_{S1}[k]} \right). \tag{25}$$

We can simplify (25) into

$$\begin{cases} \frac{\lambda_{S1} \gamma_{S1}[k]}{1 + \gamma_{S1}[k]} = \frac{\lambda_{S2} \gamma_{S2}[k]}{1 + \rho_{S1}[k] \gamma_{S1}[k] + \rho_{S2}[k] \gamma_{S2}[k]} & : \gamma_{S1}[k] < \gamma_{S2}[k] \\ \frac{\lambda_{S2} \gamma_{S2}[k]}{1 + \gamma_{S2}[k]} = \frac{\lambda_{S1} \gamma_{S1}[k]}{1 + \rho_{S1}[k] \gamma_{S1}[k] + \rho_{S2}[k] \gamma_{S2}[k]} & : \gamma_{S1}[k] > \gamma_{S2}[k]. \end{cases} \tag{26}$$

Therefore, we can obtain $\rho_{S1}[k]$ and $\rho_{S2}[k]$ using:

- If $\gamma_{S1}[k] < \gamma_{S2}[k]$ and $\lambda_{S1} \leq \lambda_{S2}$, then $\rho_{S2}[k] = 1$.
- If $\gamma_{S1}[k] > \gamma_{S2}[k]$ and $\lambda_{S1} \geq \lambda_{S2}$, then $\rho_{S1}[k] = 1$.
- If $\gamma_{S1}[k] < \gamma_{S2}[k]$ and $\lambda_{S1} > \lambda_{S2}$, then $\rho_{S1}[k]$ is given by (27),

$$\rho_{S1}[k] = \left(\frac{\lambda_{S1}\gamma_{S1}[k](1 + \gamma_{S2}[k]) - \lambda_{S2}\gamma_{S2}[k](1 + \gamma_{S1}[k])}{(\gamma_{S2}[k] - \gamma_{S1}[k])\lambda_{S1}\gamma_{S1}[k]} \right)_{L0H1} \tag{27}$$

$$\rho_{S2}[k] = \left(\frac{\lambda_{S2}\gamma_{S2}[k](1 + \gamma_{S1}[k]) - \lambda_{S1}\gamma_{S1}[k](1 + \gamma_{S2}[k])}{(\gamma_{S1}[k] - \gamma_{S2}[k])\lambda_{S2}\gamma_{S2}[k]} \right)_{L0H1} \tag{28}$$

- If $\gamma_{S1}[k] > \gamma_{S2}[k]$ and $\lambda_{S1} < \lambda_{S2}$, then $\rho_{S2}[k]$ is given by (28),

where the notation $(x)_{L0H1}$, defined as

$$(x)_{L0H1} \equiv \min\left(\max(x, 0), 1\right), \tag{29}$$

is used because $0 \leq \rho_{S1}[k] \leq 1$ and $0 \leq \rho_{S2}[k] \leq 1$. Once we obtain $\rho_{S1}[k]$ or $\rho_{S2}[k]$, we can obtain the other one since their sum equals one.

In a similar way, from (24b), and with reference to (7), we get

$$\lambda_{D1}\left(\frac{\partial R_{D1}^{ac}[k]}{\partial \rho_{D1}[k]} - \frac{\partial R_{D1}^{ac}[k]}{\partial \rho_{D2}[k]} \right) = \\ \lambda_{D2}\left(\frac{\partial R_{D2}^{ac}[k]}{\partial \rho_{D2}[k]} - \frac{\partial R_{D2}^{ac}[k]}{\partial \rho_{D1}[k]} \right). \tag{30}$$

We can simplify (30) into

$$\begin{cases} \frac{\lambda_{D1}\gamma_{D1}[k]}{1 + \rho_{D2}[k]\gamma_{D1}[k]} = \frac{\lambda_{D2}\gamma_{D2}[k]}{1 + \rho_{D2}[k]\gamma_{D2}[k]} & : \gamma_{D1}[k] < \gamma_{D2}[k] \\ \frac{\lambda_{D1}\gamma_{D1}[k]}{1 + \rho_{D1}[k]\gamma_{D1}[k]} = \frac{\lambda_{D2}\gamma_{D2}[k]}{1 + \rho_{D1}[k]\gamma_{D2}[k]} & : \gamma_{D1}[k] > \gamma_{D2}[k]. \end{cases} \tag{31}$$

Therefore, we can obtain $\rho_{D1}[k]$ and $\rho_{D2}[k]$ using:

- If $\gamma_{D1}[k] < \gamma_{D2}[k]$ and $\lambda_{D1} \leq \lambda_{D2}$, then $\rho_{D2}[k] = 1$.
- If $\gamma_{D1}[k] > \gamma_{D2}[k]$ and $\lambda_{D1} \geq \lambda_{D2}$, then $\rho_{D1}[k] = 1$.
- If $\gamma_{D1}[k] < \gamma_{D2}[k]$ and $\lambda_{D1} > \lambda_{D2}$, then

$$\rho_{D2}[k] = \left(\frac{\lambda_{D1}\gamma_{D1}[k] - \lambda_{D2}\gamma_{D2}[k]}{(\lambda_{D2} - \lambda_{D1})\gamma_{D1}[k]\gamma_{D2}[k]} \right)_{L0H1}. \tag{32}$$

- If $\gamma_{D1}[k] > \gamma_{D2}[k]$ and $\lambda_{D1} < \lambda_{D2}$, then

$$\rho_{D1}[k] = \left(\frac{\lambda_{D2}\gamma_{D2}[k] - \lambda_{D1}\gamma_{D1}[k]}{(\lambda_{D1} - \lambda_{D2})\gamma_{D2}[k]\gamma_{D1}[k]} \right)_{L0H1}. \tag{33}$$

C. Satisfying the Flow Conservation Constraints

In Section III-B we get the optimal solution of (15). The solution was independent of the channels' statistics. The remaining step is to solve (14) by obtaining λ_{S1} and λ_{S2} that satisfy the flow conservation constraints (3b) at equality. This part is dependent on the channels' statistics. We present two numerical approaches to solve this problem.

1) Bisection Search over λ_{S1} and λ_{S2} with Known Channel Statistics: For the case of orthogonal transmission, we can characterize the long-term average achievable rate of S_i ($i \in \{1, 2\}$) using

$$\bar{R}_{Si} = \int_0^\infty r f_{R_{Si}}(r) \Pr\left(S_i \text{ is scheduled} | R_{Si} = r\right) dr, \tag{34}$$

where $f_{R_{Si}}(r)$ is the probability density function (PDF)[6] of the achievable rate over the channel between S_i and the relay (i.e. $R_{Si}^{ac} = \log(1 + \gamma_{Si})$). Similarly,

$$\bar{R}_{Di} = \int_0^\infty r f_{R_{Di}}(r) \Pr\left(D_i \text{ is scheduled} | R_{Di} = r\right) dr, \tag{35}$$

With reference to (23), and with the assumption that the four links are independent, we can obtain the conditional probability term (for S_1 for example) using

$$\Pr\left(S_1 \text{ is scheduled} | R_{S1} = r\right) = \Pr\left(\lambda_{S1}r > \lambda_{S2}R_{S2}\right) \times \\ \prod_{j=1}^2 \Pr\left(\lambda_{S1}r > \lambda_{Dj}R_{Dj}\right). \tag{36}$$

Therefore, we can write

$$\Pr\left(S_i \text{ is scheduled} | R_{Si} = r\right) = \\ F_{R_{Sl}}\left(\frac{\lambda_{Si}\,r}{\lambda_{Sl}}\right) \prod_{j=1}^2 F_{R_{Dj}}\left(\frac{\lambda_{Si}\,r}{\lambda_{Dj}}\right), \quad l \in \{1, 2\}, \, l \neq i \tag{37}$$

and similarly

$$\Pr\left(D_i \text{ is scheduled} | R_{Di} = r\right) = \\ F_{R_{Dl}}\left(\frac{\lambda_{Di}\,r}{\lambda_{Dl}}\right) \prod_{j=1}^2 F_{R_{Sj}}\left(\frac{\lambda_{Di}\,r}{\lambda_{Sj}}\right), \quad l \in \{1, 2\}, \, l \neq i \tag{38}$$

where $F_R(r)$ is the cumulative distribution function (CDF)[7] of the achievable rate. The expressions for the long term rates of the ith source and ith user in (39) and (40), respectively, are applicable for all channel distributions. As an example, we give here the expressions for the special case of Rayleigh fading. For Rayleigh fading, the long term rates of the ith source and user are given by (39) and (40), respectively, where $\bar{\gamma}_{Si}$ is the average SNR of the link between the ith source and the relay and $\bar{\gamma}_{Di}$ is the average SNR of the link between the ith user and the relay.

With the aid of these analytical expressions, we can apply a two dimensional search over λ_{S1} and λ_{S2} to find the unique values that make $\bar{R}_{S1} - \bar{R}_{D1} = 0$ and $\bar{R}_{S2} - \bar{R}_{D2} = 0$.

2) Real-Time Adaptation of λ_{S1} and λ_{S2}: Notice that it is very complex to obtain analytical expressions of the average rates for the non-orthogonal transmission case. Therefore, the

---

[6]For Rayleigh fading, $f_R(r) = \frac{\ln 2}{\bar{\gamma}} 2^r \exp\left(-\frac{2^r - 1}{\bar{\gamma}}\right)$, $r \geq 0$, where $\bar{\gamma}$ is the average SNR.

[7]For Rayleigh fading, $F_R(r) = 1 - \exp\left(-\frac{2^r - 1}{\bar{\gamma}}\right)$.

$$\bar{R}_{Si} = \frac{\ln 2}{\bar{\gamma}_{Si}} \int_0^\infty r 2^r \exp\left(-\frac{2^r - 1}{\bar{\gamma}_{Si}}\right) \left(1 - \exp\left(-\frac{2^{\frac{\lambda_{Si} r}{\lambda_{Sl}}} - 1}{\bar{\gamma}_{Sl}}\right)\right) \prod_{j=1}^2 \left(1 - \exp\left(-\frac{2^{\frac{\lambda_{Si} r}{\lambda_{Dj}}} - 1}{\bar{\gamma}_{Dj}}\right)\right) dr, \qquad (39)$$

$$\bar{R}_{Di} = \frac{\ln 2}{\bar{\gamma}_{Di}} \int_0^\infty r 2^r \exp\left(-\frac{2^r - 1}{\bar{\gamma}_{Di}}\right) \left(1 - \exp\left(-\frac{2^{\frac{\lambda_{Di} r}{\lambda_{Dl}}} - 1}{\bar{\gamma}_{Dl}}\right)\right) \prod_{j=1}^2 \left(1 - \exp\left(-\frac{2^{\frac{\lambda_{Di} r}{\lambda_{Sj}}} - 1}{\bar{\gamma}_{Sj}}\right)\right) dr, \qquad (40)$$

bisection search approach cannot be applied in this case. Furthermore, in a practical deployment scenario, the ergodic PDF and CDF of the S-R and R-D channels may not be perfectly known, and hence off-line calculation of λ_{S1} and λ_{S2} might not be feasible even for the orthogonal transmission case. Therefore, we propose an alternative approach to obtain λ_{S1} and λ_{S2}, based on real-time channel measurements (or simulation). The weights are updated with each time index n according to

$$\lambda_{Si}(n) = \lambda_{Si}(n-1) + \delta\left(\bar{R}_{Di}(n) - \bar{R}_{Si}(n)\right), \qquad (41)$$

where $\bar{R}_{Si}(n)$ means the real-time average rate transmitted by S_i up to time index n. In (41), δ controls the convergence speed of the algorithm. The complete steps are given in Algorithm 1, which is run for a sufficient large number of iterations N such that we reach within a small tolerance value of the optimal values of λ_{S1} and λ_{S2}. It is noted here that Algorithm 1 belongs to the subgradient methods which are standard algorithms for convex[8] minimization and are known to converge [62].

Algorithm 1 Real-Time Adaptation of λ_{S1} and λ_{S2}

Initialize $\bar{R}_{Si}(0) = 0$, $i = 1, 2$
Initialize $\bar{R}_{Di}(0) = 0$, $i = 1, 2$
Initialize $\lambda_{S1}(0)$ and $\lambda_{S2}(0)$
Initialize $\lambda_{Di}(0) = \mu_i - \lambda_{Si}(0)$
for $n=1$ to N **do**
     Generate random channel realizations (or measure in real-time) for $R_{Si}^{ac}(n)$ and $R_{Di}^{ac}(n)$
     Obtain the optimal resource allocation using $\lambda_{Si}(n)$ and $\lambda_{Di}(n)$
     Update $\bar{R}_{Si}(n) = \bar{R}_{Si}(n-1) + \frac{1}{n}\left(R_{Si}(n) - \bar{R}_{Si}(n-1)\right)$, $i = 1, 2$
     Update $\bar{R}_{Di}(n) = \bar{R}_{Di}(n-1) + \frac{1}{n}\left(R_{Di}(n) - \bar{R}_{Di}(n-1)\right)$, $i = 1, 2$
     Update $\lambda_{S1}(n)$ and $\lambda_{S2}(n)$ using (41).
     Update $\lambda_{Di}(n) = \mu_i - \lambda_{Si}(n)$
end for
$\lambda_{Si} = \lambda_{Si}(N)$, $i = 1, 2$
return λ_{Si}, $i = 1, 2$

---

[8]The Lagrangian dual problem is a convex problem by construction [57].

IV. SOLUTION FOR RELAYING WITHOUT BUFFERING

A. Optimal Resource Allocation

We want to solve (3a) with the constraints in (4). Therefore, for each k, we want to solve

$$\max_{\xi[k]} \quad \mu_1 R_{D1}[k] + \mu_2 R_{D2}[k] \qquad (42a)$$

$$\text{subject to} \quad R_{D1}[k] \leq R_{S1}[k] \quad \text{and} \quad R_{D2}[k] \leq R_{S2}[k], \qquad (42b)$$

where the optimization variables $\xi[k]$ and the characterization of the achievable rates are similar to Section II-C.

We start by characterizing the two end points (of the achievable rate region) in which a single source-destination pair accesses the channel. Thus, $\rho_{Si} = \rho_{Di} = 1$, where $i \in \{1, 2\}$. In this case, we want to find the optimal $\tau_S[k]$. The solution yields

$$\tau_S[k] = \frac{R_{Di}^{ac}[k]}{R_{Si}^{ac}[k] + R_{Di}^{ac}[k]} = \frac{\log(1 + \gamma_{Di}[k])}{\log(1 + \gamma_{Si}[k]) + \log(1 + \gamma_{Di}[k])}, \qquad (43)$$

and the corresponding achievable rate is

$$R_{Si}[k] = R_{Di}[k] = \frac{R_{Si}^{ac}[k] R_{Di}^{ac}[k]}{R_{Si}^{ac}[k] + R_{Di}^{ac}[k]}. \qquad (44)$$

The solution of (42) for the **orthogonal transmission** case is by selecting a single source-destination pair to access the kth RU. The selected pair is the one that has the higher weighted rate ($\mu_i R_{Di}[k]$) among the two pairs, where $R_{Di}[k]$ is given by (44). Notice that for a block fading channel model with continuous PDF functions of the channels, the probability that the two pairs have exactly equal weighted rate at any RU is zero.

To obtain the solution of (42) for the **non-orthogonal transmission** case, we convert the weighted sum objective function into weighted Max-Min problem (e.g. [10]), which is *easier to solve*. Notice that the initial problem formulation in (1) involved maximizing the weighted sum of the minimum of the source rate and respective destination rate since at optimality they should be equal. Therefore, we reformulate the problem (42a) as

$$\max_{\sigma_1[k], \sigma_2[k]} \quad \mu_1 R_{D1}(\sigma_1[k], \sigma_2[k])[k] + \mu_2 R_{D2}(\sigma_1[k], \sigma_2[k])[k], \qquad (45)$$

where $\sigma_1[k]$ and $\sigma_2[k]$ are the weights of the Max-Min problem, where $\sigma_1[k] + \sigma_2[k] = 1$, and the rates as functions of

$$\max_{\xi[k]} \quad \min \quad (\sigma_1[k]R_{D1}[k], \sigma_1[k]R_{S1}[k], \sigma_2[k]R_{D2}[k], \sigma_2[k]R_{S2}[k]) \qquad (46)$$

these weights, $R_{D1}(\sigma_1[k], \sigma_2[k])[k]$ and $R_{D2}(\sigma_1[k], \sigma_2[k])[k]$, are obtained by solving (46) given on the top of the next page.

The solution of (46), must satisfy

$$\sigma_1[k]\tau_D[k]R_{D1}^{ac}[k] = \sigma_1[k]\tau_S[k]R_{S1}^{ac}[k]$$
$$= \sigma_2[k]\tau_D[k]R_{D2}^{ac}[k] = \sigma_2[k]\tau_S[k]R_{S2}^{ac}[k]. \qquad (47)$$

It is straightforward to observe from (47) that we retain the same ratio between the rates of the source and the destination for each pair, i.e. $\frac{R_{D1}^{ac}[k]}{R_{S1}^{ac}[k]} = \frac{R_{D2}^{ac}[k]}{R_{S1}^{ac}[k]}$, and hence we can satisfy the constraints in (47) by adjusting τ_S accordingly. This can be achieved by solving the following sub-problems:

- With reference to (6), find $\rho_{S1}[k]$ and $\rho_{S2}[k]$ that satisfy $\sigma_1[k]R_{S1}^{ac}[k] = \sigma_2[k]R_{S2}^{ac}[k]$. This can be obtained numerically using a one-dimensional bisection search as the problem is convex for a given $\sigma_1[k]$.
- With reference to (7), find $\rho_{D1}[k]$ and $\rho_{D2}[k]$ that satisfy $\sigma_1[k]R_{D1}^{ac}[k] = \sigma_2[k]R_{D2}^{ac}[k]$. This can also be obtained numerically using a one-dimensional bisection search as the problem is again convex for a given $\sigma_1[k]$.
- Find $\tau_S[k]$ using

$$\tau_S[k] = \frac{R_{D1}^{ac}[k]}{R_{S1}^{ac}[k] + R_{D1}^{ac}[k]} = \frac{R_{D2}^{ac}[k]}{R_{S2}^{ac}[k] + R_{D2}^{ac}[k]}, \qquad (48)$$

where R_{Si}^{ac} and R_{Di}^{ac} are obtained with the power ratios ρ_{Si} and ρ_{Di} obtained in the previous two steps.

The remaining step is to solve (45) with the aid of the solution of (46). We obtain the solution using a bisection search over $\sigma_1[k]$ to find the maximum as described in Algorithm 2, which is similar to the search method used in [41].

Algorithm 2 Bisection Search over $\sigma_1[k]$ to solve (45)

Initialize $\kappa_0 = 0$, $\kappa_2 = 0.5$, $\kappa_4 = 1$
Given a tolerance ϵ,
1) Set $\kappa_1 = (\kappa_0 + \kappa_2)/2$ and $\kappa_3 = (\kappa_2 + \kappa_4)/2$.
2) Find $m = \arg\max_{j \in \{0,1,2,3,4\}} \mu_1 R_{D1}(\kappa_j, 1 - \kappa_j) + \mu_2 R_{D2}(\kappa_j, 1 - \kappa_j)$.
3) Replace κ_0 by $\kappa_{\max(m-1,0)}$ and replace κ_4 by $\kappa_{\min(m+1,4)}$.
4) If $m \notin \{0, 4\}$, set $\kappa_2 = \kappa_m$, else set $\kappa_2 = (\kappa_0 + \kappa_4)/2$.
5) If $\kappa_4 - \kappa_0 \geq \epsilon$, return to 1), else set solution at κ_m.
return $\sigma_1[k] = \kappa_m$

B. Average Achievable Rates for Orthogonal Transmission Case

The long-term average achievable rates in the orthogonal transmission case (for $i = 1, 2$) can be obtained using

$$\bar{R}_{SDi} = \int_0^\infty r f_{R_{SDi}}(r) \times$$
$$\Pr(S_i - D_i \text{ pair is scheduled}|R_{SDi} = r)\, dr, \qquad (49)$$

where $f_{R_{SDi}}(r)$ is the PDF of the achievable rate by the S_i-D_i pair. By referring to (44), we can characterize $F_{R_{SDi}}(r)$ using two equivalent integral forms,

$$F_{R_{SDi}}(r) = \int_0^\infty f_{R_{Si}}(x) F_{R_{Di}}\left(\left[\frac{xr}{x-r}\right]^*\right) dx$$
$$= \int_0^\infty f_{R_{Di}}(x) F_{R_{Di}}\left(\left[\frac{xr}{x-r}\right]^*\right) dx, \qquad (50)$$

where we apply a similar notation to the one suggested in [63] in order to write the CDF as a single integration instead of double integration,

$$[y]^* = \begin{cases} y & : y \geq 0 \\ \infty & : y < 0. \end{cases} \qquad (51)$$

Furthermore, we can write the PDF of the achievable rate as

$$f_{R_{SDi}}(r) = \int_r^\infty \frac{x^2}{(x-r)^2} f_{R_{Si}}(x) f_{R_{Di}}\left(\frac{xr}{x-r}\right) dx$$
$$= \int_r^\infty \frac{x^2}{(x-r)^2} f_{R_{Di}}(x) f_{R_{Si}}\left(\frac{xr}{x-r}\right) dx. \qquad (52)$$

Using the defined PDF and CDF, we can characterize the conditional probability term in (49) using

$$\Pr(S_i - D_i \text{ pair is scheduled}|R_{SDi} = r) = F_{R_{SDl}}\left(\frac{\mu_i\, r}{\mu_l}\right),$$
$$l \in \{1, 2\}, \ l \neq i. \qquad (53)$$

For the case of Rayleigh fading, (49) becomes (54).

V. SUB-OPTIMAL SCHEDULING SCHEMES

In order to characterize the gains of the optimal scheduling scheme, we compare against benchmark schemes which do not exploit the diversity of the four links in the system or exploit them partially.

A. Round Robin Scheduling

We consider a scheme with no buffering at the relay and which does not take advantage of multi-user diversity and multi-hop gains. Therefore, in each RU, one source and the corresponding user will be selected in a round robin manner as

$$\bar{R}_{SDi} = \frac{(\ln 2)^2}{\bar{\gamma}_{Si}\bar{\gamma}_{Di}} \int_0^\infty r \int_r^\infty \frac{x^2}{(x-r)^2} 2^{\frac{x^2}{x-r}} \exp\left(\frac{2^x-1}{\bar{\gamma}_{Si}}\right) \exp\left(\frac{2^{\frac{xr}{x-r}}-1}{\bar{\gamma}_{Di}}\right) \left(1 - \exp\left(\frac{2^{\frac{xr}{x-r}}-1}{\bar{\gamma}_{Di}}\right)\right) dx dr \qquad (54)$$
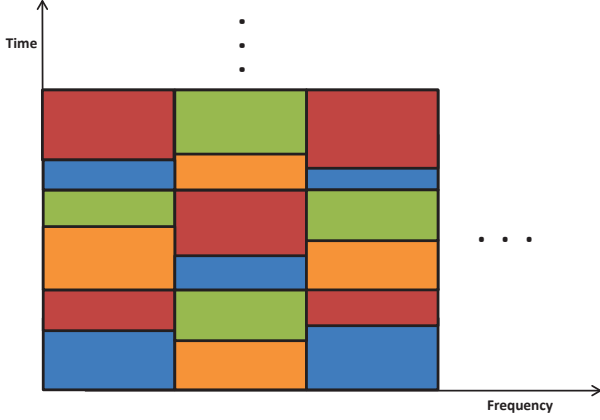

Fig. 3: Example of the allocation of RUs for round robin scheduling. Blue represents the source 1, yellow represents source 2, red represents user 1 and green represents user 2.
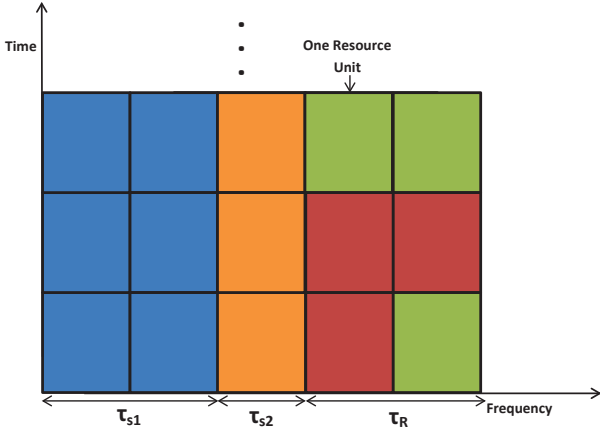

Fig. 4: Example of the allocation of RUs for multi-user scheduling only. Blue represents the source 1, yellow represents source 2, red represents user 1 and green represents user 2.

shown in Fig. 3. The corresponding $\tau_S[k]$ and the achievable rates are similar to the case discussed in (43) and (44).

Furthermore, the average achievable rate of the ith user (the same as the average rate of the ith source) is given by

$$\bar{R}_{Di} = \tau_i \int_0^\infty \int_0^\infty \frac{r_{Si} r_{Di}}{r_{Si} + r_{Di}} f_{R_{Si}}(r_{Si}) f_{R_{Di}}(r_{Di}) dr_{Si} dr_{Di},$$
$$i = 1, 2, \qquad (55)$$

where τ_i is the ratio of the RUs that are allocated to the S_i-D_i pair to the total number of RUs. For Rayleigh fading, (55) becomes (56).

B. Multi-user Scheduling only

In this scheme, which we call multi-user scheduling only, the total bandwidth is divided into three parts reserved for S_1, S_2 and the relay respectively over all time duration regardless of the channel conditions. The ratios of the total bandwidth that are allocated to each node are denoted, τ_{S1} for S_1, τ_{sS} for S_2, and τ_R for the relay. Furthermore, in the bandwidth that is reserved for the relay, the two users are scheduled opportunistically based on their channel conditions in each RU. So, the system benefits from MUD gains but not MHD gains. Fig. 4 shows an example of the allocation of RUs for the multiuser scheduling only scheme.

Notice that buffering is available at the relay since the relay does not have to forward information sent by a source to the respective user directly. The user that is selected by the relay in each RU is the one that has higher weighted rate,

$$\eta[k] = \arg\max_i \mu_i \log(1 + \gamma_{Di}[k]), \qquad (57)$$

where $i \in \{1, 2\}$.

The long-term average achievable rates of the sources are given by $\bar{R}_{Si} = \tau_{Si} \bar{R}_{Si}^{ac}$, where

$$\bar{R}_{Si}^{ac} = \int_0^\infty r f_{R_{Si}}(r) dr. \qquad (58)$$

Similarly, we can write $\bar{R}_{Di} = \tau_R \bar{R}_{Di}^{ac}$, where

$$\bar{R}_{Di}^{ac} = \int_0^\infty r f_{R_{Di}}(r) F_{R_{Dl}}\left(\frac{\mu_i r}{\mu_l}\right) dr, \quad l \neq i. \qquad (59)$$

Furthermore, from the flow conservation constraints for the relay's buffers, we have

$$\tau_{S1}\bar{R}_{S1}^{ac} = \tau_R \bar{R}_{D1}^{ac}, \ \tau_{S2}\bar{R}_{S2}^{ac} = \tau_R \bar{R}_{D2}^{ac}, \ \tau_{S1} + \tau_{S2} + \tau_R = 1. \qquad (60)$$

Solving (60) yields

$$\tau_{S1} = \frac{R_{D1}^{ac} R_{S2}^{ac}}{R_{D1}^{ac} R_{S2}^{ac} + R_{S1}^{ac} R_{S2}^{ac} + R_{D2}^{ac} R_{S1}^{ac}}, \qquad (61a)$$

$$\tau_{S2} = \frac{R_{D2}^{ac} R_{S1}^{ac}}{R_{D1}^{ac} R_{S2}^{ac} + R_{S1}^{ac} R_{S2}^{ac} + R_{D2}^{ac} R_{S1}^{ac}}, \qquad (61b)$$

$$\tau_R = \frac{R_{S1}^{ac} R_{S2}^{ac}}{R_{D1}^{ac} R_{S2}^{ac} + R_{S1}^{ac} R_{S2}^{ac} + R_{D2}^{ac} R_{S1}^{ac}}. \qquad (61c)$$

For Rayleigh fading, (58) and (59) can be written as

$$\bar{R}_{Si}^{ac} = \frac{\ln 2}{\bar{\gamma}_{Si}} \int_0^\infty r 2^r \exp\left(-\frac{2^r-1}{\bar{\gamma}_{Si}}\right) dr, \qquad (62)$$

$$\bar{R}_{Di}^{ac} = \frac{\ln 2}{\bar{\gamma}_{Di}} \int_0^\infty r 2^r \exp\left(-\frac{2^r-1}{\bar{\gamma}_{Di}}\right)$$
$$\left(1 - \exp\left(-\frac{2^{\frac{\mu_i r}{\mu_l}}-1}{\bar{\gamma}_{Dl}}\right)\right) dr, \qquad (63)$$

10

$$\bar{R}_{Di} = \tau_i \frac{(\ln 2)^2}{\bar{\gamma}_{Si}\bar{\gamma}_{Di}} \int_0^\infty \int_0^\infty \frac{r_{Si}\, r_{Di}}{r_{Si} + r_{Di}} 2^{r_{Si}} 2^{r_{Di}} \exp\left(-\frac{2^{r_{Si}}-1}{\bar{\gamma}_{Si}}\right) \exp\left(-\frac{2^{r_{Di}}-1}{\bar{\gamma}_{Di}}\right) dr_{Si}\, dr_{Di} \tag{56}$$
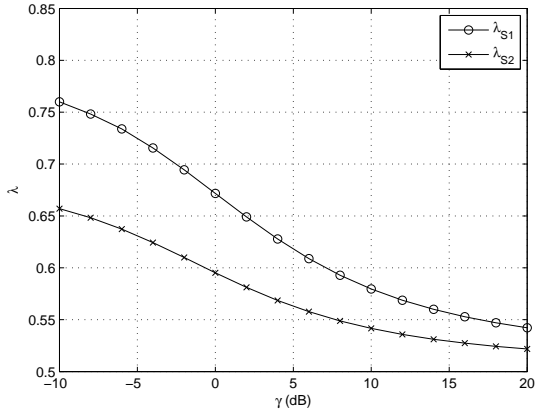

Fig. 5: Optimal λ_{S1} and λ_{S2} with $\bar{\gamma}_{S1} = 0.5\gamma$, $\bar{\gamma}_{S2} = 0.7\gamma$, $\bar{\gamma}_{D1} = \gamma$ and $\bar{\gamma}_{D2} = \gamma$.
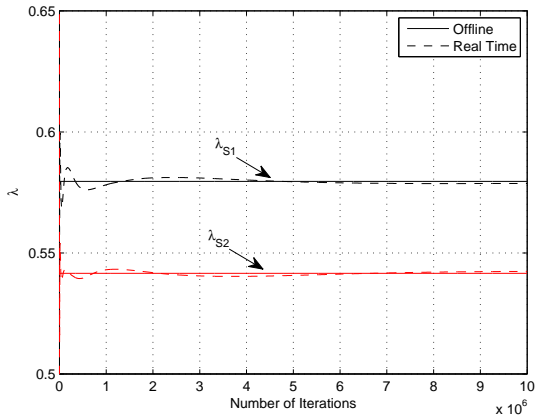

Fig. 6: Real Time update of λ with γ=10 dB.

respectively.

VI. NUMERICAL RESULTS

We present numerical results for the proposed optimal scheduling scheme as well as the different sub-optimal schemes presented in this paper. We assume that the four links of the system model are Rayleigh block-fading, where $\bar{\gamma}_{S1}$, $\bar{\gamma}_{S2}$, $\bar{\gamma}_{D1}$ and $\bar{\gamma}_{D2}$ represent the average SNR gain of the channels of S_1, S_2, D_1 and D_2, respectively. The value of K is set at 10^6.

We start by showing results related to Section III-C for obtaining the Lagrangian multipliers that satisfy the flow conservation constraints at equality. Fig. 5 shows the optimal λ_{S1} and λ_{S2} as functions of the SNRs of the channels. In this example, we assume that both destinations have the same average SNR, while S_1 has lower average SNR than S_2. We maximize the sum achievable rate in this example. Therefore, we set $\mu_1 = \mu_2 = 1$. It is evident from Fig. 5 that the optimal

λ_{Si} (for $i = 1, 2$) decreases as the SNRs increase. Also, the source with the lower average SNR has higher λ_S.

A comparison between the offline (two-dimensional search method) approach and the real-time adaptation approach is shown in Fig. 6. The recursively updated values of λ_{S1} and λ_{S2} are obtained as described in Algorithm 2. The figure demonstrates that the results of the adaptive approach converge to the optimal values obtained by the offline approach.

Next, we compare the different scheduling schemes in terms of the maximum sum of the long-term average achievable rates of the two source-destination pairs in Fig. 7, and in terms of the achievable rate region in Fig. 8. Therefore, in Fig. 7 we set $\mu_1 = \mu_2 = 1$, while in Fig. 8 we scan the achievable rates of $\bar{R}_{D1}$ and $\bar{R}_{D2}$ for values of μ_1 and μ_2 varied gradually between $(0, 1)$ and $(1, 0)$. The average channel gains are shown in the figure captions. The results in both figures demonstrate that buffering at the relay produces considerable gain in terms of the achievable rates as compared with scheduling schemes with no buffering at the relay. Furthermore, the results show that non-orthogonal transmission provides little improvement in comparison with the orthogonal transmission case for both cases of buffer-aided relaying and no buffering at the relay. The gain of non-orthogonal transmission is more negligible in the case of buffer-aided relaying. Moreover, we see from the figures that the performance of the multiuser scheduling only scheme and the schemes with no buffering at the relay are relatively close to each other. Both schemes exploit the diversity in the system partially. However, the optimal scheme exploit both MUD and MHD fully. Therefore, it produces considerable capacity gains. The round robin scheduling scheme is the worst since it does not exploit the diversity of the fading channels.

In addition, we also show the rate region obtained through simulation for the orthogonal case. In the simulation, it is assumed that the buffers are initially empty and the optimal resource allocation strategy is applied and in the event that the the second hop is scheduled, it is limited by both the buffer and the link capacity. It can be observed from Fig. 8 that the even though we started with empty buffers, the simulation results are almost the same as the analytical results which assumed that the relay is only limited by the link capacity. Hence, the assumption of being only limited by the link capacity is justified. However, as stated previously, there were a very small number of RUs in which the buffer did not support the link capacity, but their effect on the long-term average achievable rate region is evidently negligible.

VII. CONCLUSIONS

We have studied and solved the problem of optimal resource allocation over block-fading channels with the objective of maximizing the long-term average achievable rate region for dual-hop system with two source-destination pairs sharing
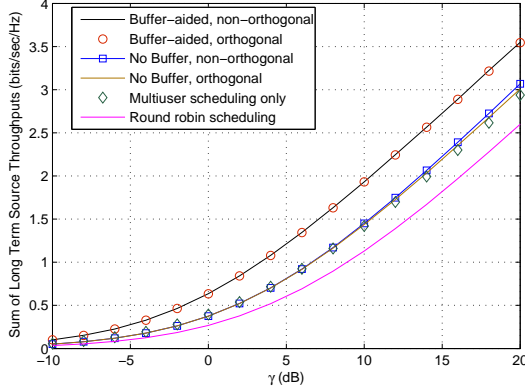
Fig. 7: Comparison of $\bar{R}_{S1}+\bar{R}_{S2}$ as a function of γ with $\bar{\gamma}_{S1} = 0.5\gamma$, $\bar{\gamma}_{S2} = 0.7\gamma$, $\bar{\gamma}_{D1} = \gamma$ and $\bar{\gamma}_{D2} = \gamma$.
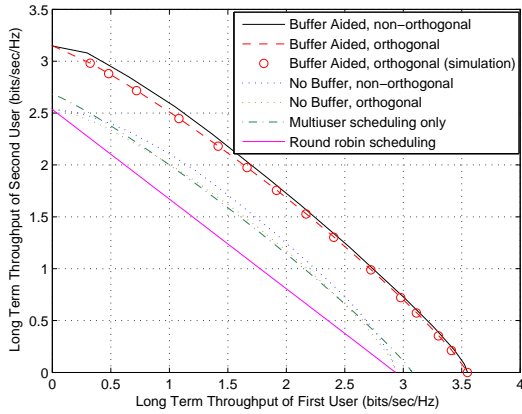

Fig. 8: Comparison of the achievable rate regions for $\gamma = 20$ dB with $\bar{\gamma}_{S1} = \gamma$, $\bar{\gamma}_{S2} = \gamma$, $\bar{\gamma}_{D1} = 1.5\gamma$ and $\bar{\gamma}_{D2} = 0.5\gamma$.

a single relay. This system model is relevant to the case when a relay assists two neighboring base stations. We have emphasized an important feature to get considerable capacity gains which is achieved by aiding the relay with buffers to enable storing the incoming data traffic temporarily until the channels of the second hop become favorable for transmission. This feature enables exploiting the multiuser diversity and the multihop diversity effects of the system.

We have shown that the optimal resource allocation scheme is joint user-and-hop scheduling in which either the sources or the relay transmit in a given resource unit and not both of them. Furthermore, under orthogonal transmission constraints, only one source transmits or only one destination receives information in a given resource unit. We have discussed how to obtain the optimal resource allocation using an offline approach based on a two-dimensional search method with known channel statistics and a real-time approach based on channel measurements or simulation.

We have provided numerical results to compare the performance of the proposed scheduling scheme with sub-optimal schemes with no buffering at the relay or with multiuser scheduling only. The numerical results demonstrated that buffering at the relay produces considerable capacity gains, and that non-orthogonal transmission provides little improve-

ment in comparison with the orthogonal transmission case. Due to its evident gains in terms of the supported transmission capacity (bits/sec/Hz), we believe that the proposed scheduling schemes for buffer-aided relays have the potential to be implemented practically in LTE-Advanced systems. To achieve this objective, we believe that more research efforts are needed to evaluate the system performance in practice and to account for some system constraints such as limited buffer sizes or restricted delay constraints.

APPENDIX A
JOINT RELAY SELECTION AND OPTIMAL RESOURCE ALLOCATION FOR MULTIPLE SHARED RELAYS

In this appendix, we consider the extension of the proposed schemes to the case of multiple relays serving two source-destination pairs. In this case, the total achievable rate at each destination will be the sum of the rates received from all available relays. Therefore, the problem in (3) can be formulated as

$$\max_{\xi_{\mathbf{v}}[k],\forall k,v} \quad \sum_{v=1}^{M} \left(\mu_1 \bar{R}_{R_v D1} + \mu_2 \bar{R}_{R_v D2} \right) \tag{64a}$$

$$\text{subject to} \quad \bar{R}_{R_v D1} \leq \bar{R}_{R_v S1} \quad \text{and} \tag{64b}$$

$$\bar{R}_{R_v D2} \leq \bar{R}_{R_v S2}, \qquad \forall v \in \{1, 2, \dots, M\}, \tag{64c}$$

where $\bar{R}_{R_v D1}$ is the long term rate from the vth relay to the first destination, $\bar{R}_{R_v D2}$ is the long term rate from the vth relay to the second destination, $\bar{R}_{R_v S1}$ is the long term rate from the first source to the vth relay, $\bar{R}_{R_v S2}$ is the long term rate from the second source to the vth relay, $v \in \{1, 2, \dots, M\}$ and M is the number of relays. Each of the M relays is equipped with two buffers for storing data, one for each source-destination pairs. Furthermore, the vector of optimization variables $\xi_v[k]$ corresponds to one relay and it has six optimization variables as defined in (8) and (11). Following the same procedure and notation as in Section III, (64) can be written as K independent problems in (65) where $\Phi(\lambda_{R_v S1}, \lambda_{R_v S2})[k]$ is given by (66), $\lambda_{R_v D1} = \mu_1 - \lambda_{R_v S1}$, $\lambda_{R_v D2} = \mu_2 - \lambda_{R_v S2}$. Furthermore, $\lambda_{R_v S1}$ and $\lambda_{R_v S2}$ ($\forall v \in \{1, 2, \dots, M\}$) are the $2M$ Lagrange multipliers associated with the flow conservation constraints in (64).

For a scheme that involves the best relay selection, it is easy to see that the optimal solution requires the best relay selection and then to decide whether to have non-orthogonal transmission from the two sources to the selected relay or non-orthogonal transmission from the selected relay to the two destinations. In this case, the relay which can support the highest weighted rate is selected. This also holds true for the case of orthogonal transmission. Thus, for multiple relays with single relay selection, the resource allocation scheme remains the same in concept. However, it becomes difficult to implement in practice even with orthogonal transmissions since we will have now $2M$ buffers, which leads to $2M$ Lagrange multipliers in the optimization problem. This will, in general, require a $2M$ dimensional search to find the optimal Lagrangian multipliers. Furthermore, for the real-time

12

$$\Phi(\lambda_{R_1S1}, \lambda_{R_2S1}, \ldots, \lambda_{R_MS1}, \lambda_{R_1S2}, \lambda_{R_2S2}, \ldots, \lambda_{R_MS2}) = \sum_{k=1}^{K} \sum_{v=1}^{M} \Phi(\lambda_{R_vS1}, \lambda_{R_vS2})[k], \quad (65)$$

$$\Phi(\lambda_{R_vS1}, \lambda_{R_vS2})[k] = \max_{\xi_{\mathbf{v}}[k]} \quad \lambda_{R_vD1}R_{R_vD1}[k] + \lambda_{R_vD2}R_{D2}[k] + \lambda_{R_vS1}R_{R_vS1}[k] + \lambda_{R_vS2}R_{R_vS2}[k], \quad (66)$$

adaptation of the weights, we need to modify Algorithm 1 to include all the Lagrange multipliers.

In summary, the extension of our resource allocation scheme to include multiple relays with single "best" relay selection is simple to characterize. However, from a practical perspective, the complexity of getting the optimal solution becomes significant as the number of relays increases. Additionally, there is another difficulty in this case to find a centralized node that can have all CSI for the links between the two sources and the two destinations to all relays in order to make the resource allocation decisions accordingly.

Furthermore, for multiple relays systems in which all the relays are used in an all-participate setting, then it is even more difficult to find the optimal transmission strategy since the two sources transmit to different relays in this case. Similarly, there can be transmissions from a number of relays to the two destinations. Optimizing the rate and channel allocation based on all the channel information of all relays is a very difficult and open problem. This is beyond the scope of this paper.

REFERENCES

[1] 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.

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

[3] S. Sesia, I. Toufik, and M. Baker, *LTE – The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons, Ltd., 2009.

[4] Proceedings of the IEEE, *Special Issue on Adaptive Modulation and Transmission in Wireless Systems*, vol. 95, no. 12, December 2007.

[5] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[6] D. Tse and S. Hanly, "Multiaccess fading channels – Part 1: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[7] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels – Part 1: Ergodic capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

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

[9] M. Shaqfeh, H. Alnuweiri, and M.-S. Alouini, "Multiuser switched diversity scheduling schemes," *IEEE Transactions on Communications*, vol. 60, no. 9, pp. 2499–2510, Sep. 2012.

[10] 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.

[11] 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.

[12] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "Joint scheduling for dual-hop block-fading broadcast channels," in *Proc. 8th IEEE Broadband Wireless Access Workshop (BWA'2012) in conjunction with IEEE Global Communications Conference (GlobeCom'2012)*, Anaheim, CA, Dec. 2012, pp. 140–145.

[13] IEEE Journal on Selected Areas in Communications, *Special Issue on Cooperative Communications and Networking*, vol. 25, no. 2, February 2007.

[14] IEEE Transactions on Wireless Communications, *Special Issue on Cooperative Communications*, vol. 7, no. 5, May 2008.

[15] 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.

[16] 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.

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

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

[19] ——, "A survey of multi-way channels in information theory: 1961-1976," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 1–37, Jan. 1977.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] M. Hasna and M. S. Alouini, "Optimal power allocation for relayed transmissions over Rayleigh-fading channels," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1999–2004, Nov. 2004.

[26] ——, "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.

[27] ——, "End-to-end performance of transmission systems with relays over Rayleigh-fading channels," *IEEE Transactions on Wireless Communications*, vol. 2, no. 6, pp. 1126–1131, Nov. 2003.

[28] ——, "Harmonic mean and end-to-end performance of transmission systems with relays," *IEEE Transactions on Communications*, vol. 52, no. 1, pp. 130–135, Jan. 2004.

[29] 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.

[30] G. Kramer and A. van Wijngaarden, "On the white Gaussian multiple-access relay channel," in *Proc. IEEE International Symposium on Information Theory (ISIT'2000)*, Sorrento, Italy, Jun. 2000, p. 40.

[31] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, Sept. 2005.

[32] L. Sankaranarayanan, G. Kramer, and N. Mandayam, "Capacity theorems for the multiple-access relay channel," in *Proc. 42nd Annual Allerton Conference on Communications, Control, and Computing 2004*, Monticello, Illinois, Sep. 2004, pp. 1782–1791.

[33] I. Maric and R. Yates, "Cooperative multicast for maximum network lifetime," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 127–135, Jan. 2005.

[34] ——, "Bandwidth and power allocation for cooperative strategies in Gaussian relay networks," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1880–1889, Apr. 2010.

[35] 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.

[36] D. Brown III, "Resource allocation for cooperative transmission in wireless networks with orthogonal users," in *Proc. 38th Asilomar Conference on Signals, Systems and Computers 2004*, vol. 2, Pacific Grove, CA, Nov. 2004, pp. 1473–1477.

[37] A. Scaglione and Y. Hong, "Opportunistic large arrays: Cooperative transmission in wireless multihop ad hoc networks to reach far distances," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2082–2092, Aug. 2003.

[38] A. Host-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, Jun. 2005.

[39] L. Sankaranarayanan, G. Kramer, and N. Mandayam, "Cooperation vs. hierarchy: An information-theoretic comparison," in *Proc. IEEE International Symposium on Information Theory (ISIT'2005)*, Adelaide, Australia, Sep. 2005, pp. 411–415.

[40] S. Serbetli and A. Yener, "Relay assisted F/TDMA ad hoc networks: Node classification, power allocation and relaying strategies," *IEEE Transactions on Communications*, vol. 56, no. 6, pp. 937–947, 2008.

[41] 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.

[42] B. Xia, Y. Fan, J. Thompson, and H. V. Poor, "Buffering in a three-node relay network," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4492–4496, Nov. 2008.

[43] N. Zlatanov, R. Schober, and P. Popovski, "Buffer-aided relaying with adaptive link selection," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 8, pp. 1530–1542, Aug. 2013.

[44] I. Krikidis, T. Charalambous, and J. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1957–1967, May 2012.

[45] A. Ikhlef, D. Michalopoulos, and R. Schober, "Max-max relay selection for relays with buffers," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1124–1135, Mar. 2012.

[46] A. Ikhlef, J. Kim, and R. Schober, "Mimicking full-duplex relaying using half-duplex relays with buffers," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 7, pp. 3025–3037, Jul. 2012.

[47] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "Exploiting multi-user diversity and multi-hop diversity in dual-hop broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3314–3325, Jul. 2013.

[48] 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.

[49] 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.

[50] A. E. Gamal, M. Mohseni, and S. Zahedi, "Bounds on capacity and minimum energy-per-bit for AWGN relay channels," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1545–1561, Apr. 2006.

[51] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 2, pp. 359–378, 1994.

[52] G. Kaplan and S. Shamai, "Achievable performance over the correlated rician channel," *IEEE Transactions on Communications*, vol. 42, no. 11, pp. 2967–2978, 1994.

[53] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, 1998.

[54] M. Shaqfeh and N. Goertz, "Performance analysis of scheduling policies for delay-tolerant applications in centralized wireless networks," in *Proc. IEEE International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'08)*, Edinburgh, UK, Jun. 2008, pp. 309–316.

[55] 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.

[56] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

[57] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, USA: Cambridge University Press, 2009.

[58] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[59] M. Shaqfeh, W. Mesbah, and H. Alnuweiri, "Utility maximization for layered transmission using the broadcast approach," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1228–1238, Mar. 2012.

[60] W. Mesbah, M. Shaqfeh, and H. Alnuweiri, "Jointly optimal rate and power allocation for multilayer transmission," *IEEE Transactions on Wireless Communications*, accepted for publication, to appear in 2014.

[61] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "On multiple users scheduling using superposition coding over Rayleigh fading channels," *IEEE Communications Letters*, vol. 17, no. 4, Apr. 2013.

[62] S. Boyd and A. Mutapcic, "Stochastic subgradient methods," Class notes for EE364b, Winter 2006-07, available at:http://see.stanford.edu/materials/lsocoee364b/04-stoch_subgrad_notes.pdf.

[63] 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.

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 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.