“© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”
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 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

![System Model: Dual-hop system with two source-destination pairs.](image)

We consider the system shown in Fig. 1, where two sources (S₁ and S₂) send information to two users (destinations D₁ and D₂) through a single shared relay (R). In this system, S₁ sends information to D₁ only and S₂ sends information to D₂ 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 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, R₁[1] and R₂[1], 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.

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 (D₁ and D₂) 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) weights² for the two users. We can formulate the optimization problem as

\[
\max \left( \mu₁ \min(\bar{R}_₁, \bar{R}_₂) + \mu₂ \min(\bar{R}_₂, \bar{R}_₂) \right),
\]

where \( \mu₁ \) and \( \mu₂ \) are given weights. Also, \( \bar{R}_₁ \) and \( \bar{R}_₂ \) are the average rates transmitted by S₁ and S₂, respectively, and decoded reliably at the relay. Similarly, \( \bar{R}_₁ \) and \( \bar{R}_₂ \) are the average rates transmitted by the relay and decoded reliably at D₁ and D₂, respectively. The averaging is over all RUs. We give index \( k \) for the resource units. Therefore, we can write

\[
\bar{R}_₁ = \frac{1}{K} \sum_{k=1}^{K} R₁[₁],
\]

where \( R₁[₁] \) is the rate transmitted by S₁ in the \( k \)th RU, and \( K \) is the total number of RUs, which is assumed to be very large. The same notation can be applied to \( \bar{R}_₂, \bar{R}_₁ \) and \( \bar{R}_₂ \).

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 \( \xi[₁] \) to denote the vector of optimization variables for the \( k \)th RU.

The minimum of \( \bar{R}_₁ \) and \( \bar{R}_₂ \) is taken in (1) since the information sent by S₁ to D₁ goes over dual hops and hence the rate is bounded by the worst hop. This is also the case for

²As 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.
where the long-term average input rate to the queue of the buffer should be greater than or equal to the long-term average output rate. However, the optimal solution of (3) is achieved when the constraint is satisfied at equality.

Moreover, we must distribute the resources such that \( R_{S1} = \bar{R}_{D1} \) and \( R_{S2} = \bar{R}_{D2} \). 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

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

subject to \( \bar{R}_{D1} \leq \bar{R}_{S1} \) and \( \bar{R}_{D2} \leq \bar{R}_{S2} \), \( \forall k \) \)

where \( k \) stands for \( k = \{1,2,\ldots,K\} \). 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. 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,

\[
R_{D1}[k] = R_{S1}[k] \quad \text{and} \quad R_{D2}[k] = R_{S2}[k] \quad \forall k. \quad (4)
\]

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 \( k \)th RU to the total time of the RU are denoted \( \tau_S[k] \) and \( \tau_D[k] \), respectively. Therefore, we can write

\[
R_{S1}[k] = \tau_S[k] R_{S1}^{ac}[k] \quad (5a)
\]

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

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 \( k \)th RU. Similarly, \( R_{D1}^{ac}[k] \) is the achievable rate of the channel between the relay and \( D_1 \) in the \( k \)th 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 \( k \)th 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. 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. 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

\[
R_{S1}^{ac}[k] = \left\{ \begin{array}{ll}
\log \left( 1 + \rho_{S1}[k] \gamma_{S1}[k] \right) & : \gamma_{S1}[k] < \gamma_{S2}[k] \\
\log \left( 1 + \frac{\rho_{S1}[k] \gamma_{S1}[k]}{\tau_S[k] \gamma_{S2}[k]} \right) & : \gamma_{S1}[k] > \gamma_{S2}[k]
\end{array} \right. \quad (6)
\]

\[
R_{D1}^{ac}[k] = \left\{ \begin{array}{ll}
\log \left( 1 + \frac{\rho_{D1}[k] \gamma_{D1}[k]}{\tau_D[k] \gamma_{D2}[k]} \right) & : \gamma_{D1}[k] < \gamma_{D2}[k] \\
\log \left( 1 + \frac{\rho_{D1}[k] \gamma_{D1}[k]}{\tau_D[k] \gamma_{D2}[k]} \right) & : \gamma_{D1}[k] > \gamma_{D2}[k].
\end{array} \right. \quad (7)
\]

where \( \gamma \) 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

\[
\xi[k] = \begin{bmatrix} \tau_S[k], \tau_D[k], \rho_{S1}[k], \rho_{S2}[k], \rho_{D1}[k], \rho_{D2}[k] \end{bmatrix}. \quad (8)
\]

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

\[
\tau_S[k] + \tau_D[k] = 1, \quad \tau_S[k] > 0, \quad \tau_D[k] > 0, \quad (9a)
\]

\[
\rho_{S1}[k] + \rho_{S2}[k] = 1, \quad \rho_{S1}[k] \geq 0, \quad \rho_{S2}[k] \geq 0, \quad (9b)
\]

\[
\rho_{D1}[k] + \rho_{D2}[k] = 1, \quad \rho_{D1}[k] \geq 0, \quad \rho_{D2}[k] \geq 0. \quad (9c)
\]

When orthogonal transmission is used, we have

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

The capacity region of the multiple-access channel with sum power constraint is the same as 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

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

The space of the vector $\xi[k]$ in this case is defined by $\omega_{S1}[k] + \omega_{S2}[k] = 1, \omega_{S1}[k] \geq 0, \omega_{S2}[k] \geq 0$, (12a) $\omega_{D1}[k] + \omega_{D2}[k] = 1, \omega_{D1}[k] \geq 0, \omega_{D2}[k] \geq 0$, (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 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

$$R_{D1} = \frac{1}{K} \sum_{k=1}^{K} \tau_D[k] \min (Q_1[k], R_{D1}^{ac}[k]) = \frac{1}{K} \sum_{k=1}^{K} \tau_D[k] R_{D1}^{ac}[k],$$

(13)

and the same applies to $R_{D2}$.

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

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

(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 $k$th RU are only dependent on their respective optimization variables in the $k$th RU (i.e. $\xi[k]$) and are independent of the optimization variables in all other RUs. Therefore, (16) is equivalent to

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

(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 $\lambda_{S1}$ and $\lambda_{S2}$. 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 $\lambda_{S1}$ and $\lambda_{S2}$ 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

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

(21a)

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

(21b)

Notice that we should replace $\rho$ by $\omega$ in (21) for the orthogonal transmission case. Furthermore, the optimal solution of $\tau_S[k]$ in (18) is
\[
\Phi(\lambda_{S1}, \lambda_{S2}) = \max_{\xi[k], \nu[k]} \mu_1 R_{D1} + \mu_2 R_{D2} - \lambda_{S1}(R_{D1} - \bar{R}_{S1}) - \lambda_{S2}(R_{D2} - \bar{R}_{S2})
\]

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

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

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

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

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

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 \( \omega[k] \), where

\[
\omega[k] = \arg \max_{x} \lambda_x \log(1 + \gamma_x[k]),
\]

(23)

where \( x \in \{S_1, S_2, D_1, D_2\} \).

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

\[
\frac{\partial \Phi_S[k]}{\partial \rho_{S1}[k]} = \frac{\partial \Phi_S[k]}{\partial \rho_{S2}[k]} = 0, \quad \rho_{S1}[k] + \rho_{S2}[k] = 1,
\]

\[
\frac{\partial \Phi_D[k]}{\partial \rho_{D1}[k]} = \frac{\partial \Phi_D[k]}{\partial \rho_{D2}[k]} = 0, \quad \rho_{D1}[k] + \rho_{D2}[k] = 1.
\]

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

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

We can simplify (25) into

\[
\begin{align*}
\frac{\lambda_{S1} \gamma_{S1}[k]}{1 + \gamma_{S1}[k]} & = \frac{\lambda_{S2} \gamma_{S2}[k]}{1 + \gamma_{S2}[k]}, & \gamma_{S1}[k] < \gamma_{S2}[k] \\
\frac{\lambda_{S1} \gamma_{S1}[k]}{1 + \gamma_{S1}[k]} & = \frac{\lambda_{S2} \gamma_{S2}[k]}{1 + \gamma_{S2}[k]}, & \gamma_{S1}[k] > \gamma_{S2}[k].
\end{align*}
\]

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).

\(^5\text{It is well-known that with non-orthogonal transmission, we can achieve higher capacity gains by allowing multiple (sources / users) scheduling per RU as demonstrated numerically in [61] for a broadcast channel case.}\)
For the case of orthogonal transmission, we can satisfy the flow conservation constraints (3b) at equality. This 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
\[
\rho_{D1} \left( \frac{\partial R_{ac}^{\rho \rho_{D2}}[k]}{\partial \rho_{D1}[k]} \right) = \frac{\partial R_{ac}^{\rho \rho_{D2}}[k]}{\partial \rho_{D2}[k]},
\]
\[
\rho_{D2} \left( \frac{\partial R_{ac}^{\rho \rho_{D2}}[k]}{\partial \rho_{D1}[k]} \right).
\]

We can simplify (30) into
\[
\frac{\lambda_{D1}\gamma_{D1}[k]}{1+\rho_{D2}[k]_\gamma_{D1}[k]} = \frac{\lambda_{D2}\gamma_{D2}[k]}{1+\rho_{D2}[k]_\gamma_{D2}[k]} : \gamma_{D1}[k] < \gamma_{D2}[k],
\]
\[
\frac{\lambda_{D1}\gamma_{D1}[k]}{1+\rho_{D2}[k]_\gamma_{D1}[k]} = \frac{\lambda_{D2}\gamma_{D2}[k]}{1+\rho_{D2}[k]_\gamma_{D2}[k]} : \gamma_{D1}[k] \geq \gamma_{D2}[k].
\]

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

- If \gamma_{D1}[k] \geq \gamma_{D2}[k] and \lambda_{D1} \geq \lambda_{D2}, then \rho_{D2}[k] = 0.
- If \gamma_{D1}[k] \geq \gamma_{D2}[k] and \lambda_{D1} \leq \lambda_{D2}, then \rho_{D2}[k] = 1.
- If \gamma_{D1}[k] > \gamma_{D2}[k] and \lambda_{D1} \leq \lambda_{D2}, then \rho_{D2}[k] = \left( \frac{\lambda_{D1}\gamma_{D1}[k] - \lambda_{D2}\gamma_{D2}[k]}{(\lambda_{D1} - \lambda_{D2})\gamma_{D1}[k]\gamma_{D2}[k]} \right)_{LOH1}.
- If \gamma_{D1}[k] > \gamma_{D2}[k] and \lambda_{D1} \geq \lambda_{D2}, then \rho_{D2}[k] = \left( \frac{\lambda_{D2}\gamma_{D2}[k] - \lambda_{D1}\gamma_{D1}[k]}{(\lambda_{D1} - \lambda_{D2})\gamma_{D1}[k]\gamma_{D2}[k]} \right)_{LOH1}.

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 \lambda_{S1} and \lambda_{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 \lambda_{S1} and \lambda_{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
\[
\bar{R}_{Si} = \int_0^\infty r f_{R_{Si}}(r) P_r (S_i \text{ is scheduled}\mid R_{Si} = r) \, dr,
\]

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. \( R_{ac}^{\gamma} = \log(1 + \gamma_{S_i}) \)). Similarly,
\[
\bar{R}_{Di} = \int_0^\infty r f_{R_{Di}}(r) P_r (D_i \text{ is scheduled}\mid R_{Di} = r) \, dr,
\]

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
\[
P_r (S_1 \text{ is scheduled}\mid R_{S1} = r) = P_r (\lambda_{S1}r > \lambda_{S2}R_{S2}) \times \prod_{j=1}^2 P_r (\lambda_{S1}r > \lambda_{Dj}R_{Dj}).
\]

Therefore, we can write
\[
P_r (S_1 \text{ is scheduled}\mid R_{S1} = r) = F_{R_{S1}} \left( \frac{\lambda_{S1}r}{\lambda_{S1}} \right) \prod_{j=1}^2 F_{R_{Dj}} \left( \frac{\lambda_{S1}r}{\lambda_{Dj}} \right), \quad l \in \{1, 2\}, l \neq i
\]

and similarly
\[
P_r (D_i \text{ is scheduled}\mid R_{Di} = r) = F_{R_{Di}} \left( \frac{\lambda_{D1}r}{\lambda_{Di}} \right) \prod_{j=1}^2 F_{R_{Si}} \left( \frac{\lambda_{D1}r}{\lambda_{Sj}} \right), \quad l \in \{1, 2\}, l \neq i
\]

where \( F_{R}(r) \) is the cumulative distribution function (CDF)\(^7\) of the achievable rate. The expressions for the long term rates of the \( i \)-th source and \( i \)-th 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 \( i \)-th source and user are given by (39) and (40), respectively, where \( \bar{\gamma}_{S1} \) is the average SNR of the link between the \( i \)-th source and the relay, and \( \bar{\gamma}_{Di} \) is the average SNR of the link between the \( i \)-th user and the relay.

With the aid of these analytical expressions, we can apply a two dimensional search over \( \lambda_{S1} \) and \( \lambda_{S2} \) 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 \( \lambda_{S1} \) and \( \lambda_{S2} \): 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, \( f_{R}(r) = \frac{\ln 2}{r} 2^r \exp \left( -\frac{r-\bar{\gamma}}{\bar{\gamma}} \right), \quad r \geq 0, \) where \( \bar{\gamma} \) is the average SNR.

\(^{7}\)For Rayleigh fading, \( F_{R}(r) = 1 - \exp \left( -\frac{r-\bar{\gamma}}{\bar{\gamma}} \right) \).
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 \( \lambda_{S1} \) and \( \lambda_{S2} \) might not be feasible even for the orthogonal transmission case. Therefore, we propose an alternative approach to obtain \( \lambda_{S1} \) and \( \lambda_{S2} \), based on real-time channel measurements (or simulation). The weights are updated with each time index \( n \) according to

\[
\lambda_{S(i)}(n) = \lambda_{S(i)}(n-1) + \delta (\bar{R}_{D(i)}(n) - \bar{R}_{S(i)}(n)) ,
\]

where \( \bar{R}_{S(i)}(n) \) means the real-time average rate transmitted by \( S_i \) up to time index \( n \). In (41), \( \delta \) 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 \( \lambda_{S1} \) and \( \lambda_{S2} \). 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 \( \lambda_{S1} \) and \( \lambda_{S2} \)

1. Initialize \( \bar{R}_{S(i)}(0) = 0, \quad i = 1, 2 \)
2. Initialize \( \bar{R}_{D(i)}(0) = 0, \quad i = 1, 2 \)
3. Initialize \( \lambda_{S(0)} \) and \( \lambda_{S(0)} \)
4. Initialize \( \lambda_{D(0)} = \mu_i - \lambda_{S(i)}(0) \)
5. for \( n = 1 \) to \( N \) do
   1. Generate random channel realizations (or measure in real-time) for \( R_{S(i)}^u(n) \) and \( R_{D(i)}^u(n) \)
   2. Obtain the optimal resource allocation using \( \lambda_{S(i)}(n) \) and \( \lambda_{D(i)}(n) \)
   3. Update \( \bar{R}_{S(i)}(n) = \bar{R}_{S(i)}(n-1) + \frac{1}{\delta} (R_{S(i)}(n) - \bar{R}_{S(i)}(n-1)) \), \( i = 1, 2 \)
   4. Update \( \bar{R}_{D(i)}(n) = \bar{R}_{D(i)}(n-1) + \frac{1}{\delta} (R_{D(i)}(n) - \bar{R}_{D(i)}(n-1)) \), \( i = 1, 2 \)
   5. Update \( \lambda_{S(i)}(n) \) and \( \lambda_{S(i)}(n) \) using (41).
   6. Update \( \lambda_{D(i)}(n) = \mu_i - \lambda_{S(i)}(n) \)
6. end for

Return \( \lambda_{S(i)}(N), \quad i = 1, 2 \)

---

**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

\[
\max_{\xi[k]} \mu_1 R_{D1}[k] + \mu_2 R_{D2}[k]
\]

subject to \( R_{D1}[k] \leq R_{S1}[k] \) and \( R_{D2}[k] \leq R_{S2}[k] \),

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

\[
\tau_S[k] = \frac{R_{D1}^c[k]}{R_{S1}^c[k] + R_{D1}^c[k]} = \frac{\log(1 + \gamma_{D1}[k])}{\log(1 + \gamma_{S1}[k]) + \log(1 + \gamma_{D1}[k])},
\]

and the corresponding achievable rate is

\[
R_{Si}[k] = R_{Di}[k] = \frac{R_{S1}^c[k] R_{D1}^c[k]}{R_{S1}^c[k] + R_{D1}^c[k]}.
\]

The solution of (42) for the orthogonal transmission case is by selecting a single source-destination pair to access the \( k \)th RU. The selected pair is the one that has the higher weighted rate (\( \mu_i R_{Di}[k] \)) among the two pairs, where \( \bar{R}_{Di}[k] \) 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

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

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

\[\text{The Lagrangian dual problem is a convex problem by construction [57].}\]
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]) \quad (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
\[
\begin{align*}
\sigma_1[k]T_D[k]R_{D1}^{ac}[k] &= \sigma_1[k]T_S[k]R_{S1}^{ac}[k] \\
&= \sigma_2[k]T_D[k]R_{D2}^{ac}[k] = \sigma_2[k]T_S[k]R_{S2}^{ac}[k].
\end{align*}
\]
(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., \( \frac{R_{S1}^{ac}[k]}{R_{S2}^{ac}[k]} = \frac{R_{D1}^{ac}[k]}{R_{D2}^{ac}[k]} \), and hence we can satisfy the constraints in (47) by adjusting \( \tau_S \) 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
\[
\tau_S[k] = \frac{R_{D1}^{ac}[k]}{R_{S1}^{ac}[k] + R_{D1}^{ac}[k]} = \frac{R_{D2}^{ac}[k]}{R_{S2}^{ac}[k] + R_{D2}^{ac}[k]},
\]
(48)
where \( R_{S1}^{ac} \) and \( R_{D1}^{ac} \) are obtained with the power ratios \( \rho_{S1} \) and \( \rho_{D1} \) 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 \( \epsilon \),

1. Set \( \kappa_1 = (\kappa_0 + \kappa_2)/2 \) and \( \kappa_3 = (\kappa_2 + \kappa_4)/2 \).
2. Find \( m = \max_{j \in \{0,1,2,3,4\}} \mu_1 R_{D1}(\kappa_j, 1 - \kappa_j) + \mu_2 R_{D2}(\kappa_j, 1 - \kappa_j) \).
3. Replace \( \kappa_0 \) by \( \kappa_{\max(m-1,0)} \) and replace \( \kappa_4 \) by \( \kappa_{\min(m+1,4)} \).
4. If \( m \notin \{0,4\} \), set \( \kappa_2 = \kappa_0 \), else set \( \kappa_2 = (\kappa_0 + \kappa_4)/2 \).
5. If \( \kappa_4 - \kappa_0 \geq \epsilon \), return 1, else set solution at \( \kappa_m \).

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
\[
\bar{R}_{SDi} = \int_0^\infty r f_{R_{SI}}(r) dr,
\]
(49)
where \( f_{R_{SI}}(r) \) 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,
\[
F_{R_{SDi}}(r) = \int_0^\infty f_{R_{SI}}(x) F_{R_{Di}}  \left( \left[ \frac{xt}{x-r} \right] \right) dx
\]
(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,
\[
[y]^* = \begin{cases} y: & y \geq 0 \\ \infty: & y < 0 \end{cases}
\]
(51)
Furthermore, we can write the PDF of the achievable rate as
\[
f_{R_{SDi}}(r) = \int_r^\infty \frac{x^2}{(x-r)^2} f_{R_{SI}}(x) f_{R_{Di}}  \left( \frac{xt}{x-r} \right) dx
\]
(52)
Using the defined PDF and CDF, we can characterize the conditional probability term in (49) using
\[
\Pr (S_i - D_i \text{ pair is scheduled } | R_{SDi} = r) = F_{R_{SDi}}  \left( \frac{\mu_i r}{\mu} \right),
\]
(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
\[ R_{SD_i} = \frac{(\ln 2)^2}{\gamma_{Si}^2 \gamma_{Di}} \int_0^{\infty} r \int_0^\infty \frac{x^2}{(x-r)^2} 2 \frac{\gamma}{\gamma_{Si}} \exp \left( \frac{2x - 1}{\gamma_{Si}} \right) \exp \left( \frac{2x - 1}{\gamma_{Di}} \right) \left( 1 - \exp \left( \frac{2x - 1}{\gamma_{Di}} \right) \right) dx dr \]  

(54)

**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, \( \tau_{S1} \) for \( S_1 \), \( \tau_{SS} \) for \( S_2 \), and \( \tau_R \) 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,

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

(57)

where \( i \in \{1, 2\} \).

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

\[ \bar{R}_{Si} = \int_0^{\infty} r f_{R_{Si}}(r) dr. \]  

(58)

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

\[ \bar{R}_{Di} = \int_0^{\infty} r f_{R_{Di}}(r) F_{R_{Di}}(\mu_i r \mu_i) dr, \quad l \neq i. \]  

(59)

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

\[ \tau_{S1} \bar{R}_{S1} = \tau_R \bar{R}_{D1}, \quad \tau_{S2} \bar{R}_{S2} = \tau_R \bar{R}_{D2}, \quad \tau_{S1} + \tau_{S2} + \tau_R = 1. \]  

(60)

Solving (60) yields

\[ \tau_{S1} = \frac{R_{ac}}{\bar{R}_{D1}}, \quad \tau_{S2} = \frac{R_{ac}}{\bar{R}_{D2}}, \quad \tau_R = \frac{R_{S1} + R_{S2}}{\bar{R}_{D1} + \bar{R}_{D2}}. \]  

(61a)

(61b)

(61c)

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

\[ \bar{R}_{ac} = \frac{1}{\gamma_{Di}} \int_0^{\infty} r^2 \exp \left( \frac{-2x - 1}{\gamma_{Di}} \right) dr, \]  

(62)

\[ \bar{R}_{ac} = \frac{1}{\gamma_{Di}} \int_0^{\infty} r^2 \exp \left( \frac{-2x - 1}{\gamma_{Di}} \right) \left( 1 - \exp \left( \frac{-2x - 1}{\gamma_{Di}} \right) \right) dr, \]  

(63)
\[
\bar{R}_{Di} = \frac{\tau_i (\ln 2)^2}{\gamma_{Si} \gamma_{D_i}} \int_0^\infty \int_0^\infty \frac{r_{Si} r_{D_i}}{r_{Si} + r_{D_i}} 2^{r_{Si} r_{D_i}} \exp \left(-\frac{2^{r_{Si}} - 1}{\gamma_{Si}}\right) \exp \left(-\frac{2^{r_{D_i}} - 1}{\gamma_{D_i}}\right) dr_{Si} dr_{D_i}
\]

(56)

\[\lambda_{S1} \text{ (for } i = 1, 2) \text{ decreases as the SNRs increase. Also, the source with the lower average SNR has higher } \lambda_S.\]

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 \(\lambda_{S1}\) and \(\lambda_{S2}\) 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 \(\mu_1\) and \(\mu_2\) 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.

**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 \(\lambda_{S1}\) and \(\lambda_{S2}\) 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

\[\lambda_{S1} \text{ and } \lambda_{S2} \text{ are shown in Fig. 5.} \]

respectively.

![Graph showing optimal λ₁ and λ₂ with γ=10 dB](image)

**Fig. 5: Optimal λ₁ and λ₂ with γ₁ = 0.5γ, γ₂ = 0.7γ, γ₁₂ = γ and γ₁₂ = γ.**

![Graph showing real-time update of λ](image)

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

**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
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 improvement 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

\[
\max_{\xi_{v}[k],\gamma_{v},v} \sum_{v=1}^{M} (\mu_{1} R_{R,D1} + \mu_{2} R_{R,D2})
\]

subject to

\[
R_{R,D1} \leq R_{R,S1} \quad \text{and} \quad R_{R,D2} \leq R_{R,S2}, \quad \forall v \in \{1,2,\ldots,M\},
\]

where \(R_{R,D1}\) is the long term rate from the \(v\)th relay to the first destination, \(R_{R,D2}\) is the long term rate from the \(v\)th relay to the second destination, \(R_{R,S1}\) is the long term rate from the first source to the \(v\)th relay, \(R_{R,S2}\) is the long term rate from the second source to the \(v\)th relay, \(v \in \{1,2,\ldots,M\}\) 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 pair. 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 \(\Phi(\lambda_{R,S1},\lambda_{R,S2})[k]\) is given by (66), \(\lambda_{R,D1} = \mu_{1} - \lambda_{R,S1}, \lambda_{R,D2} = \mu_{2} - \lambda_{R,S2}\). Furthermore, \(\lambda_{R,S1}\) and \(\lambda_{R,S2} (\forall v \in \{1,2,\ldots,M\})\) 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
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 is another difficulty in this case to find a centralized node that can include all the Lagrange multipliers.

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


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.