Throughput–Guaranteed Resource Allocation Algorithms for Relay–aided Cellular OFDMA System

Megumi Kaneko*#, *IEEE Member*, Petar Popovski*, *IEEE Member*, Kazunori Hayashi#, *IEEE Member*
* Department of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark
#Graduate School of Informatics, Kyoto University, Yoshida Honmachi Sakyo–ku, Kyoto, 606-8501, Japan
Email: {mek|petarp}@es.aau.dk, kazunori@i.kyoto-u.ac.jp

***Abstract*— Radio resource allocation for the Downlink (DL) transmissions in a cellular system, based on Orthogonal Frequency Division Multiple Access (OFDMA), has been subject of many research studies over the past years. Nowadays, increasing attention is turned upon cellular system with relays, but only few algorithms have been designed for OFDMA based relay system. In this work, resource allocation schemes are proposed for the cases of one and multiple Relay Stations (RS) in the cell. Due to the specific design of these algorithms, which operate in a RS–aided BS centralized manner, the amount of required Channel State Information (CSI) and algorithm complexity are minimized, making them suited for practical use. The idea of adaptive RS activation is introduced, where the frame structure is adapted depending on the active RS. Different number of relays are considered and the corresponding algorithms are adapted accordingly. The simulation results show that our algorithms achieve a very good throughput/outage trade–off.**

***Index Terms*— Multi–Carrier System, Orthogonal Frequency Division Multiple Access (OFDMA), Relay System, Radio Resource Allocation, Multi–User Diversity**

I. Introduction

The demand for ubiquitous, high data rate wireless services has been ever more increasing during these recent years. One key challenge for the upcoming 4^{th} Generation ($4G$) wireless system is to enable a ubiquitous high data rate coverage. For the physical layer, Orthogonal Frequency Division Multiplexing (OFDM) transmission technology is a very promising candidate due to its high achievable spectral efficiency by bit loading and inherent robustness against inter–symbol interference caused by frequency–selective fading. In a cellular system, the difficulty for providing high data rate coverage lies in the fundamental fact that, for a fixed power and bandwidth, increasing the data rate will necessarily decrease the cell coverage. With the traditional cellular architecture, providing high data rate coverage would require deployment of a large number of Base Stations (BS) to cover the whole area, which is very costly. That is the reason why recently, there has been a great interest in the concept of relaying. By installing fixed relays in strategic positions in a cell, higher data rates can be provided in remote or shadowed areas of the cell. Above all, fixed relays are low–cost devices which can be deployed easily [1]. Hence, the problem of resource allocation and scheduling for relay–aided cellular systems has been a flourishing topic for investigation and has motivated a number of works such as [2] [3] [4]. However, these works do not consider physical layer based on OFDM. In parallel, a large research effort was directed towards the design of resource allocation algorithms for Orthogonal Frequency Division Multiple Access (OFDMA) system without relays. There is a special interest in designing resource allocation schemes in OFDMA system as it offers the possibility to exploit multi–user diversity gain [5] by scheduling each user on subcarriers where he experiences high channel gain, which can maximize cell throughput, such as in [6]. Among many others, [7] and [8] considered minimum user rate requirements; in [9], [10], [11], the Proportional Fair Scheduler (PFS), implemented in Qualcomms' HDR system for single–carrier [12], was extended to a multi–carrier system.

However, these existing algorithms for cellular OFDMA system without relays are not directly applicable to the case with relays: a specific allocation is needed for the Relay Station (RS), since a RS is neither source nor sink for the communication traffic, while the radio resource allocation for the RS can be done either by the BS or by the RS itself. This is a difficult problem with many degrees of freedom. To perform the optimal allocation, the BS should gather all the Channel State Information (CSI) of all the users for all links. However, this translates into a huge amount of signalling and algorithm complexity. Thus, practical algorithms for such a system are needed. We have proposed such algorithms when there is one relay in [13], and for multiple relays in [14]. However, the algorithms in [13] are only designed to maximize the system throughput, and [14] considers only a specific case of multiple relays in the cell. In this paper, we extend these initial studies in several ways: designing PFS–based algorithms for the single relay case, consideration of the general case of multiple relays, as well as a broader discussion that completely covers our approach. Note that the main point of introducing relays is to provide service to the users which are not efficiently served by the BS, thereby decreasing the system outage. Otherwise, had the goal been to maximize the system throughput, then the users closest to the BS should be served directly by the BS, for example using the Max CSI algorithm which allocates the user with the best CSI. Thus, our main goal is to design

A part of this work has been presented at the 2007 IEEE VTC–Spring conference, Dublin, Ireland and at the 2007 IEEE ICC conference, Glasgow, Scotland

practical allocation schemes for an OFDMA based system with relays, that not only decrease the outage, but also provide a good throughput performance, while keeping the complexity low and minimizing the required amount of CSI. The paper is outlined as follows. After presenting the system model and path selection method used in this work, the problem of resource allocation for the single RS case is considered and algorithms for subchannel/time allocation are proposed. In the second part, the algorithms are generalized to the multiple RS case, where different numbers of relays are considered. The CSI reduction obtained with our proposed algorithms is shown. The algorithms are then evaluated and compared with reference algorithms. Finally, the conclusions of this work are drawn and future directions are suggested.

II. System Model

We focus on the Downlink (DL) transmissions from a BS to Mobile Stations (MS) or RS in a single cell, where users feed back to the BS their CSI on every subchannel, defined as a group of adjacent subcarriers. The relays always decode the packets and remodulate them before forwarding them to the MSs. We use a Discrete Adaptive Modulation (DAM) model where the rates $[1, 2, 4, 6, 8]$[b/s/Hz] corresponding to the modulations [BPSK,QPSK,16-QAM,64-QAM,256-QAM] are supported when the link Signal–to–Noise–Ratio (SNR) is above their respective predefined SNR thresholds $[-5, 13.6, 20.6, 26.8, 32.9]$ [dB]. The SNR thresholds from QPSK to 256-QAM are determined for a target Bit Error Rate (BER) of 10^{-6} for uncoded M–QAM symbols in flat fading channels with a known fixed gain, based on [15] [16]. For SNRs below 13.6dB, BPSK is used, and there is no transmission below -5dB. The modulation is adapted for each user and each subchannel. Each RS is half–duplex, e. g., it can not transmit on one channel and receive on another at the same time. We assume fixed relays which can be deployed so that BS–RS is in a Line–of–Sight (LOS) condition, as in [1] [17]. This ensures a high channel quality on average over time/frequency. Thus, in the allocation, one crucial assumption is that all the BS–RS subchannels for one relay are allocated with the same rate, corresponding to the average SNR over the subchannels. The algorithms are designed to work independently in single–cell environment, since the goal of this work is to design algorithms for allocating OFDMA resource with relays in a cell, based on CSI reporting. Within the cell, the interference between simultaneously transmitting relays is considered. However, we will also evaluate the performance in a multi–cell environment with a simple interference model.

A. Frame structure

In the single RS case, the system is modelled as an axis between the BS and a RS, along which users are generated. The RS is placed at a distance of $0.8 \times R_c$ from the BS, where R_c is the cell radius. The frame structure for the single RS case is shown in Fig. 1, where the total frame length is denoted T_F. We take the following assumptions,

- The transmissions between BS and RS are divided in time by T_{BS} and T_{RS} respectively. The portion of the frame with BS–originated transmissions is referred to as *BS–subframe*, and the portion with RS–originated transmissions, *RS–subframe*.
- Inside the BS–subframe, BS–MS (e. g., direct) transmissions, and BS–RS (e. g., feeder) transmissions are allocated different subchannels.
- T_{BS}, T_{RS} can be adapted per frame, the basic assumption being equal time division.

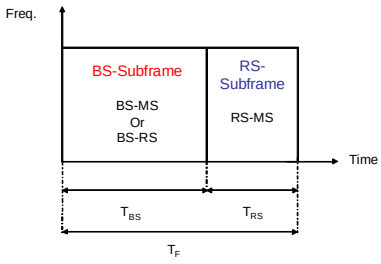

Fig. 1. Frame structure, case of single relay

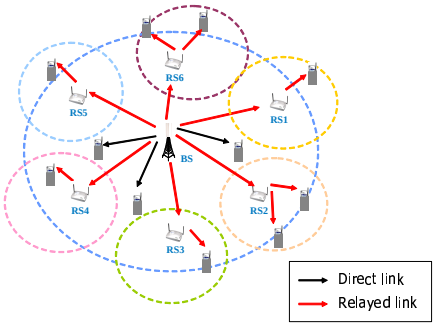

Fig. 2. Cellular System with Multiple Relays

The algorithms are optimized for the two–hop scenario. With this frame structure, the packets of a relayed user queued at the BS require at least two frames to be received: in the first frame packets are sent from BS to RS; in the second frame, from RS to MS. That is, packets sent to the RS in a frame can not be immediately forwarded due to hardware limitations, since the RS needs to store and then process the newly incoming packets. Such an operation is commonly adopted in the emerging standards that use relay–based extension of the cellular systems [18] [19]. But even in the case where the RS can send the packets in the same frame, the basic idea of our algorithms can be applied straightforwardly.

In the multiple RS case, the BS is surrounded by I equidistant RS, as depicted in Fig. 2 for $I = 6$. In [14], two types of frame division between the different relays were considered, e. g., either in time or in frequency. It was shown that both frame structures gave a similar performance. Thus, here the study will be limited to the time division case, where each RS is served sequentially in time. Since the interference between the diametrally opposed relays in the cell, e. g., the pairs

(RS_1, RS_4), (RS_2, RS_5) and (RS_3, RS_6) for $I = 6$ can be assumed low, the structure in Fig. 3 will be considered, where the frequencies are reused between 2 opposite RS.

B. Path Selection

The resource allocation essentially includes two degrees of freedom: the path selection, where a user is attached either to the direct or a relayed link, second, the subchannel/time allocation. The optimal allocation requires joint optimization of these two degrees of freedom by the BS. Since our goal is to provide low complexity algorithms with a reduced amount of CSI, we consider that the path selection is performed first, based on the long–term average user SNR, followed by the resource allocation. Since the data for relayed users takes at least two frames to be delivered, whereas the data for direct users takes only one frame, a user is linked to the RS only if $\bar{r}_k^{RS-MS} \geq 2 \times \bar{r}_k^{BS-MS}$, and to the BS otherwise, where $\bar{r}_k^l$ denotes the achievable DAM rate averaged over time and frequency for user k on link l. In the case of multiple relays, the direct rate is compared with the rate of the best relay, for that user. The long–term channel quality for each user is periodically probed, and the path selection is renewed if there is a change. Clearly, one may argue that the average rate is not a good estimate as the user is usually allocated to a channel which is better than the average. However, such a more precise estimation is quite complex, while on the other hand the adopted estimation method offers a good decision, as illustrated through the comparison with the upper bound algorithms (discussed in Section V).

III. Case of Single Relay

The target of this work is to design low complexity algorithms with good throughput and outage performance, while minimizing the required CSI. To achieve these goals, different approaches have been introduced. First, the allocations of RS–subframe and BS-subframe are decoupled: the relay makes its own tentative allocation of the RS–subframe and then, informs the BS about it. After that the BS performs the allocation of the BS–subframe and conducts optimization by considering the whole frame. This is a *RS–aided centralized* mechanism that enables a complexity decrease, since each subframe allocation is performed separately, but also a significant CSI reduction, as it will be shown in section VI. Moreover, the packets to be sent on the BS–RS link are strategically chosen: it is the relay which selects the users for which packets should be forwarded from the BS and sends this request to the BS. The selected users are the ones which have the best scheduling metric (proportional to the channel quality) but do not have any packet in the relay queue. Since low mobility users are considered, it can be reasonably assumed that a high channel quality for a certain user is likely to be kept for the next frame. Thus, with such a "just–in–time" request for user packets, it is highly probable that these packets will be delivered to the destination in the following frame. The idea behind this is that, if the BS forwarded all the packets for the relayed users as they arrive in the BS queue, there would be less resource available for the direct users, as the BS–subframe is shared between the direct users and the BS–RS link. This situation becomes critical in the heavily loaded case. In other words, to avoid penalizing the direct users, we optimize the number of subchannels allocated to the BS–RS link, by scheduling only the packets which will be surely delivered over the following frames. The effectiveness of this scheme will be shown later in section III-C.

Details of the proposed two algorithms are presented below. In the first one, the subchannel allocation is made with the equal time division, $T_{BS} = T_{RS} = T_F/2$. The optimization of the time division is performed by an iterative approach in the second algorithm.

A. Fixed Time Division (FTD) Algorithm

1. Allocation of RS–subframe by RS: In each subchannel n, relayed users are sorted in the order of best $\phi_{k,n}$, where

$$\phi_{k,n} = \frac{r_{k,n}}{\frac{\bar{\beta}_k(t-1)}{R}}, \tag{1}$$

where R is the minimum data rate requirement. $\bar{\beta}_k(t-1)$ is the past average rate allocated to user k up to frame $t-1$ over an averaging time window of p frames and is updated after every frame allocation (as in Proportional Fair Scheduling (PFS)),

$$\bar{\beta}_k(t) = \frac{p-1}{p} \times \bar{\beta}_k(t-1) + \frac{1}{p} \times \sum_{n=1}^{N} c_{k,n} r_{k,n}(t), \tag{2}$$

$c_{k,n}$ is equal to one if subchannel n was allocated to user k and zero otherwise, so $\sum_{n=1}^{N} c_{k,n} r_{k,n}(t)$ is the sum of allocated rates to user k in the current frame t. The ϕ–metric is very similar to the PFS metric, but the average allocated rates are weighted by each user's rate requirement. While in PFS, it is understood that the required rate is equal to the average allocated rate, ϕ introduces an additional degree of freedom by differentiating the required rate from the average allocated rate. That is, users whose allocated average rates are higher than their required rate are penalized, while users whose allocated rates are low compared to their required rate are prioritized, thereby decreasing the outage probability. At the same time, as in PFS, users experiencing higher instantaneous CSI are prioritized, which increases the achieved throughput. The algorithms using ϕ will be referred to as *throughput–guaranteed* algorithms, since they strive to allocate to each user a rate that best matches R. The user with highest $\phi_{k,n}$ and with packets queued at RS is allocated n.

The users having a higher $\phi_{k,n}$ than the allocated one but without packets queued at RS are represented by the set U_{Req}. For these users, the RS requests the BS to send their packets in the BS–subframe.

2. RS sends request message to BS: This message contains

- the IDs of the users in U_{Req} for which packets are requested,
- the order of these users in terms of ϕ–metric, e. g., the maximum $\phi_{k,n}$ over all n for user k in U_{Req} (needed to determine the priority of the packets sent on the BS–RS link, see Step 3)

- the value of $\bar{\phi}_{max}$, defined as the maximum value of the ϕ–metric for the relayed users in U_{Req}, averaged over the subchannels (see Step 3).

3. Allocation of BS–subframe made by BS, based on RS request: BS allocates (tentatively) each subchannel to the best direct user. The final allocation involves the allocation of BS–RS subchannels, and is conducted as follows. If U_{Req} is non empty, the scheduler calculates the number of subchannels n_{BR} required to send all the packets queued at the BS of the users in U_{Req}. As mentioned in section II, all the BS–RS subchannels are allocated with the same rate, corresponding to their average SNR level. Thus, any n_{BR} subchannels among all N subchannels can be chosen for the BS–RS transmission. To ensure a fair distribution of subchannels between the direct users and the BS–RS links, a criteria based on ϕ–metrics is introduced as follows: for each user in U_{Req}, the average ϕ–metric over the RS–MS subchannels is determined, and the maximum average ϕ–metric is denoted $\bar{\phi}_{max}$. In each subchannel, the ϕ–metric of the initially allocated direct user is compared with $\bar{\phi}_{max}$, and the subchannel is allocated to the link with the highest value. This gives y subchannels allocated to the BS–RS link. But not all y subchannels may be required, so we compare y with n_{BR},

1) If $y < n_{BR}$, the y subchannels are not enough for all the packets, e. g., some remain at the BS queue. To decide which packets to send on the BS–RS link, the RS–MS users for which packets were required are ordered by ϕ (only this order needs to be fed back to BS, not the ϕ values nor the CSI since these are not needed at the BS). Packets are allocated from the best RS–MS users, until all y subchannels are filled.
2) If $y > n_{BR}$, all y subchannels are not needed for the BS–RS link since there are less queued packets. Only the n_{BR} worst subchannels for direct users are allocated to the BS–RS link, and the remaining $y - n_{BR}$ subchannels to the best direct users.

For the FTD algorithm where $T_{BS} = T_{RS} = T_F/2$, the cell throughput $\tau_{T_F/2}$ is determined as a function of the allocated user rates and allocated packets. We define the channel utilization metric $u^l_{k,n}$ as

$$u^l_{k,n} = \frac{\min(r^l_{k,n} \times T^l_{k,n}, q_{k,n})}{r^l_{k,n} \times T^l_{k,n}}, \tag{3}$$

where l is either "BM" for direct link or "RM" for relayed link. $T^l_{k,n}$ is the number of time slots allocated to user k on subchannel n. With the equal time division, we have here $T^l_{k,n} = T_F/T_{slot}/2$ for all k, all n, where T_{slot} denotes the duration of the minimum time allocation unit. $r^l_{k,n} \times T^l_{k,n}$ is the capacity of user k on subchannel n (in number of packets, with an adequate packet size), and $q_{k,n}$ is the number of allocated packets for user k on subchannel n. Simply, if there are enough packets to fill the whole subchannel, then $u_{k,n} = 1$, otherwise all the queued packets are allocated and $u_{k,n} < 1$ since there are less packets than the available capacity. Note that $u_{k,n}$ is time dependent since the number of allocated packets depends on the time allocated. Thus, the throughput achieved in the BS–subframe by the direct users $k \in D$ is written as

$$\begin{aligned}\tau_{BM}(T_{BS} = T_F/2) &= \sum_{k \in D} \tau^k_{BM}(T_{BS} = T_F/2) \\ &= \frac{1}{T_{BS}} \sum_{k \in D} c^{BM}_{k,n} \times u^{BM}_{k,n} \times r^{BM}_{k,n}\end{aligned} \tag{4}$$

where $c^{BM}_{k,n}$ is equal to 1 if user k is allocated on subchannel n in the BS–subframe and 0 otherwise. By applying the same formula for the throughput achieved in the RS–subframe $\tau_{RM}(T_{RS} = T_F/2)$, the overall throughput is written as

$$\begin{aligned}\tau_{T_F/2} = &\frac{\tau_{BM}(T_{BS} = T_F/2) \times T_F/2}{T_F} \\ &+ \frac{\tau_{RM}(T_{RS} = T_F/2) \times T_F/2}{T_F}.\end{aligned} \tag{5}$$

The throughput achieved by the feeder link (BS–RS) is not accounted for, since the data is not delivered to the users.

B. Adaptive Time Division (ATD) Algorithm

Starting from the allocation by the FTD algorithm for $T_{BS} = T_{RS} = T_F/2$, the time division can be adapted in order to increase the overall throughput. Basically, the goal of this algorithm is to balance the allocated time between BS and RS–subframes, according to the amount of packets in the BS and RS queues. By matching the time division with the proportion of the packets in each queue, the throughput should be increased. We perform an iterative optimization to find the best time division between the BS and RS subframes, within the fixed total frame length T_F. For this optimization, the users and the number of initially allocated packets are known for each subchannel, as the outcome of the FTD algorithm. But when the time division is changed, the number of packets allocated to each user changes: for example, if the BS–subframe is increased by one slot and the RS–subframe reduced by one, the direct users are allocated an amount of packets corresponding to the additional capacity in their allocated subchannel (if there are any queued packets), and for the relayed users, the corresponding packets are removed. This optimization problem can be formulated as follows:

$$\begin{array}{ll} & \tau_{Opt} = \max_x \tau_x \\ \text{where} & \tau_x = \frac{\tau_{BM}(T_{BS}) \times T_{BS} + \tau_{RM}(T_{RS}) \times T_{RS}}{T_F} \\ \text{subject to} & T_{BS} = T_F/2 + x \times T_{slot}, T_{RS} = T_F/2 - x \times T_{slot} \\ & x \in [-\frac{T_F/2}{T_{slot}}, ..., +\frac{T_F/2}{T_{slot}}] \end{array} \tag{6}$$

where the variable $x \in \mathbf{Z}$. Since this a difficult problem due to the discrete packet updates at each time adaptation, we propose the following *Adaptive Time Division* (ATD) algorithm which performs an iterative optimization. The idea is to start from the initial condition with $T_{BS} = T_{RS} = T_F/2$ and then consider the two possible cases:

1) increase T_{BS} by one slot and decrease T_{RS} by one slot, e. g., $T_{BS} = T_F/2 + T_{slot}$ and $T_{RS} = T_F/2 - T_{slot}$. The new throughput, after the updated packet allocation,

becomes:

$$\tau_a = \frac{\tau_{BM}(T_F/2+T_{slot})\times(T_F/2+T_{slot})}{T_F} + \frac{\tau_{RM}(T_F/2-T_{slot})\times(T_F/2-T_{slot})}{T_F}. \quad (7)$$

2) decrease T_{BS} by one slot and increase T_{RS} by one slot, and the throughput becomes:

$$\tau_b = \frac{\tau_{BM}(T_F/2-T_{slot})\times(T_F/2-T_{slot})}{T_F} + \frac{\tau_{RM}(T_F/2+T_{slot})\times(T_F/2+T_{slot})}{T_F}. \quad (8)$$

Then, we compare $\tau_{T_F/2}$, τ_a and τ_b and the maximum determines the direction for the time adaptation: if $\tau_{T_F/2}$ is the maximum, the algorithm stops since the adaptation in either direction gives a lower throughput, so $\tau_{Opt} = \tau_{T_F/2}$. Otherwise, for example if the maximum is τ_a, we adapt again by increasing the BS–subframe by one slot, $T_{BS} = T_F/2 + 2\times T_{slot}$ and $T_{RS} = T_F/2 - 2\times T_{slot}$, and determine the throughput denoted τ_{a+1}. At iteration i, we have $T_{BS} = T_F/2 + i\times T_{slot}$ and $T_{RS} = T_F/2 - i\times T_{slot}$, with throughput τ_{a+i}. The iterative search stops when: $\tau_{a+i+1} < \tau_{a+i}$. We obtain the maximum throughput $\tau_{Opt} = \tau_{a+i}$.

The optimal solution requires a full search over all the possible time division. This ATD algorithm is suboptimal since it restricts the search to either one direction. However, it can still be ensured that a good division is found when the search is restrained to the direction giving an increased throughput, while decreasing the algorithm complexity. This is due to the fact that the fluctuations of the throughput in each subframe is governed by the number of packets at the queues at the BS and the relay. That is, if by adding one slot in one direction, the throughput becomes lower, it is likely that it will continue to decrease if more slots are added, as there are no more packets in the queue that can fill up the additionally allocated capacity. At the same time, the number of slots initially allocated to the other subframe are removed, which implies that the packets that were allocated initially are removed accordingly, which leads to an overall decrease of throughput. Thus, this creates a situation where one subframe is attributed too much capacity which is not utilized due to a lack of queued packets, and the other subframe has too little capacity and can not allocate a sufficient number of packets. We can observe that the final decision is made by the BS, so an overall optimization can be achieved up to a certain degree. Moreover, it will be shown in section VI that the amount of CSI feedback is reduced compared to the optimal BS–centralized algorithm. However, the BS needs to know the queue status at the RS for this adaptation, which is possible (the BS can monitor the queue status at the RS since it knows how many packets were sent to the RS and how many were allocated from the RS to the MS), but makes the ATD algorithm less practical than the FTD algorithm. In the simulations, the ATD algorithm is used to assess the FTD algorithm.

C. Discussion on the packet requests to the BS by the relay

In the FTD and ATD algorithms, the relay requests the BS to forward packets for some chosen relayed users, e. g., with the best ϕ–metric but without packets at the RS queue. This is a more complex way as compared to the case where the BS forwards all the packets for relayed users, e. g., in the order of arrival. For comparison, we have designed the *All Forward* (All–Fwd) algorithm which works like the FTD algorithm, except that the RS no longer requests packets but the BS forwards the packets for randomly chosen relayed users. That is, not all the packets for relayed users are always forwarded; in the BS–subframe, only the subchannels where the maximum of the average ϕ–metric of all relayed users, $\bar{\phi}_{max}$, is higher than the best direct user's ϕ–metric, $\phi_{k,n}$, are allocated to the BS–RS link, which determines the number of relayed users' packets that can be accommodated. The simulations will show the effectiveness of our requesting scheme compared to the *All Forward* algorithm, in terms of throughput and system outage probability (see section VII). This is the reason why the algorithms for multiple relays presented in the following section are also based on this requesting mechanism.

IV. Case of Multiple Relays

We present algorithms designed for multiple relays in the cell. Among the various algorithms in [14], we only retain the best ones. It was shown that the algorithms operating with a fixed frame were inefficient. Thus, we proposed the idea of *Adaptive RS–Activation*, where only the relays which can increase the performance measures are activated, by reallocating an under–utilized relay subframe to another entity than can use it more efficiently. The algorithms using this mechanism, *Multiple–RS Parallel with Activation* (MRPA) and *Multiple–RS Adaptive Activation* (MRAA) algorithms, enabled a very good performance enhancement. In this work, these algorithms are further adapted to varying number of relays in the cell.

A. Multiple–RS Parallel with Activation (MRPA) Algorithm

If I is an even number, we can make $I/2$ relay pairs by regrouping the diametrally opposed relays. Each group of 2 relays transmit at the same time by sharing the same RS_j–subframe, $j \in [1..I/2]$, since the interference is minimized between the opposite relays. After the path selection is performed following the procedure in section II-B, there may be relays which are *activated* or *not activated*, e. g., relays where users are attached and others where no users are attached. If there happens to be 2 opposite relays that are not activated (for example, RS_1 and RS_4), their corresponding subframe is removed and the overall RS–subframe is redivided in time among the remaining groups. As shown in Fig. 3, for $I = 6$, each subframe has an initial length of $T_F/6$ but after removal of a group, it will become $T_F/4$. However, for $I = 3$ the relays cannot be paired so frequency reuse is not considered. Instead, each of the 3 equally divided RS_j–subframe is only used by one relay. If no users are attached to a relay after the path selection, the corresponding subframe is removed and the whole RS–subframe is divided among the remaining relays. This step will be referred to as *long–term RS activation* step.

After this step, the subchannel allocation is based on the *FTD* algorithm from section III-A. The difference from the single RS case is that the algorithm is performed by each RS,

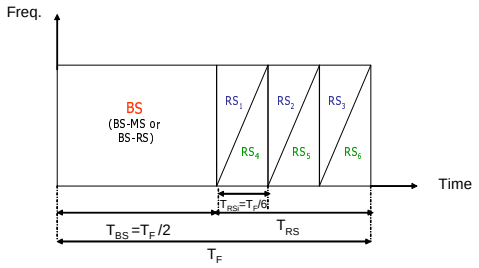

Fig. 3. Frame Structure for the Multiple–RS Parallel with Activation (MRPA) Algorithm

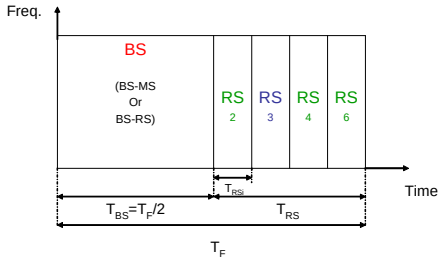

Fig. 4. Frame Structure for the Multiple–RS Adaptive Activation (MRAA) Algorithm

for each RS_j–subframe, and each RS makes its own list of users, $U_{Req}^{RS_i}$, $i \in [1..I]$, for which packets are requested. Then, the BS considers the request lists of all relays and performs the steps in III-A: in each subframe, each subchannel is allocated to the user with the best $\phi_{k,n}$, defined in Eq. (1). For the allocation of subchannels to the BS–RS_i links, the packets for the users in $U_{Req}^{RS_i}, \forall i$ are concatenated and allocated together. Each relay decodes which packets it should forward as they know the ID of their attached users. Thus, the relays only need to know which subchannels are allocated to the BS–RS links, so the signalling remains the same as in the single RS case. However, with frequency reuse for $I > 3$, some subchannels are used at the same time, so the opposite relays transmitting in parallel interfere with each other.

B. Multiple–RS Adaptive Activation (MRAA) Algorithm

For this algorithm, we consider an initial frame where $T_{BS} = T_F/2$ and the RS–subframe is equally divided in time between all the relays, without assuming frequency reuse, resulting in a RS_j–subframe length of $T_F/2/I$. The best adaptation of the time division between the different entities would be to perform an *ATD*–like algorithm as in the single RS case, where the time given to each subframe would be optimized. However, in the multiple RS case, this becomes a multi–variable integer optimization problem, which requires a very high complexity. Therefore, we propose the following method, where the RS activation occurs in two steps: the long–term RS activation and the per–frame RS activation.

The long–term RS activation works as described above, based on the path selected by each user. The idea is to keep only the RS_j–subframe for the relays where users have been attached. For example with $I = 6$, if nobody is attached to RS_1 and RS_5, the remaining access points are the BS and the other four RS. The RS–subframe duration is equally divided among the four RS, resulting in the frame shown in Fig. 4. This is a *long–term* RS activation as it is based on the path selection which depends on the long–term average channel qualities. Then, the same algorithm for subchannel allocation as in the *MRPA* algorithm is performed, based on $\phi_{k,n}$. After this initial allocation, the users and packets on each subchannel and subframe are known. Next, the throughput is optimized by removing the worst RS_j–subframes until the best throughput is obtained. This *per–frame RS activation* phase works as follows. The initial throughput τ_0 is calculated for all the R_0 relays and the BS. The BS–subframe has an initial length of $T_{BS}(0) = T_f/2$ and each RS_j–subframe has a length of $T_{RS_i} = T_f/2/R_0$. Then, the RS are sorted in the order of decreasing throughput τ_{RS_i} achieved in each RS_j–subframe. The packet allocation and the queues are updated accordingly. The RS with the worst throughput is removed, and the RS_j–subframe is reallocated to the BS–subframe, so that: $R_1 = R_0 - 1$ and $T_{BS}(1) = T_{BS}(0) + T_{RS_i}$. If the new total throughput τ_i is higher than the previous one τ_{i-1}, we continue by removing the next worst RS and redistributing the frame, otherwise the previous frame configuration is kept. This algorithm stops when $\tau_{i-1} > \tau_i$ or when only the BS subframe is left. A concern might be that a RS is always removed if this RS always achieves a low throughput, but since all RS are statistically identical, they have an equal chance to be served. As the simulation shows, this problem does not arise.

For a high number of relays, as for $I = 12$, each RS_j–subframe length $T_F/2/I$ becomes very small. If T_F/I becomes smaller than the minimum time allocation unit of one slot T_{slot}, I_{max} relays are randomly chosen to be allocated in this frame, where $I_{max} = T_F/2/T_{slot}$, and the other $I - I_{max}$ relays are discarded for this frame. In each frame, a new set of I_{max} relays are randomly chosen for allocation. After this additional step, the algorithm works as described above.

V. Design of Optimized Algorithms for Performance Assessment

To assess the performance of the proposed algorithms, optimized algorithms have been designed to provide some upper/lower bounds. So far the path selection and the subchannel/time allocation have been separated in order to reduce the algorithm complexity and to avoid the feedback of the CSI of relayed users to the BS. However, to obtain the best performance, the path selection and the resource allocation should be jointly made at the BS.

A. Upper Bound for Throughput in the case of Single Relay

The following assumptions are taken to ensure an upper bound to the throughput:

1) for a relayed user, all the packets coming from the BS to the RS in a frame are received by the user during the same frame,

2) the time division between BS–subframe and RS–subframe is optimized *per subchannel*.

In this case, the frame becomes as in Fig. 5, where $T_{k,n}^{BS}$ is the time allocated for BS–RS transmission for user k on subchannel n. For direct users, $T_{k,n}^{BS} = T_F$. This algorithm gives an optimized performance since we consider unrealistic assumptions: in reality, packets for relayed users need at least 2 frames to arrive to destination, and a different time division per subchannel is not feasible since a RS cannot receive and transmit at the same time on different subchannels. The real throughput upper bound would be given by taking the full buffer assumption (there are always packets to be sent for all users), and performing the Max CSI algorithm. But that results in an extremely disparate upper bound as the best user always achieves the highest rate due to the user distribution on the RS–BS axis. To obtain a tighter bound, this assumption was dropped and the following optimized scheme was adopted:

1) We consider K users. Set D_{CSI} contains the CSI of all K users for the direct link and set R_{CSI} contains the CSI of all K users for the relayed link. For each subchannel, we have to determine which user on which link to allocate, in order to maximize the throughput.
2) For the users in R_{CSI}, we determine the time division between $T_{k,n}^{RS}$ and $T_{k,n}^{BS}$. With the assumption that everything sent from BS to RS arrives at the MS during the same frame, $T_{k,n}^{RS}$ and $T_{k,n}^{BS}$ are proportional to the BS–RS and RS–MS rates on the subchannel, namely $\bar{r}_{BR}$ and $r_{k,n}^{RM}$, and can be determined as:

$$T_{k,n}^{RS} = \frac{\bar{r}_{BR}}{\bar{r}_{BR} + r_{k,n}^{RM}} \times T_F \tag{9}$$

and $T_{k,n}^{BS} = T_F - T_{k,n}^{RS}$.
3) The effective capacity $\eta_{k,n}^{l}$, is defined as a product of the capacity (see section III-A) with the channel utilization metric defined in (3), for each user and subchannel

$$\eta_{k,n}^{l} = u_{k,n}^{l} \times r_{k,n}^{l} \times T_{k,n}^{l}. \tag{10}$$

4) In each subchannel, we simultaneously order by decreasing effective capacities $\eta_{k,n}^{l}$ the $2K$ users from sets D_{CSI} and R_{CSI}, in order to choose the best path and the highest efficiency simultaneously. The best user, who has either link BS or RS, is allocated the subchannel. Finally, the throughput is computed.

This scheme is referred to as *RS–Max Optimized* algorithm.

B. Upper Bound for Throughput in the case of Multiple Relays

In the case of multiple RS, compared to the previous case, the same number of users is generated over a much larger area, thereby decreasing the probability that one user will always support the highest rate 8 [b/s/Hz]. Since the throughput achieved by the Max CSI algorithm assuming full buffer is much lower than in the single RS case, this algorithm gives a suitable upper bound, referred as *RS–Max Full Buffer*. As in section V-A, we also assumed that the data sent to a RS is immediately forwarded to the user. The best user among all the direct or relayed users is scheduled per subchannel.

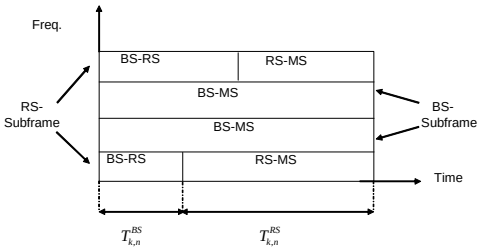

Fig. 5. Frame structure used for the optimal algorithm

C. Lower Bound for System Outage

In the case of system outage, all algorithms are evaluated with the full buffer assumption. That is, a user is considered to be in outage when he has not been given any resource, although he has queued packets. But if he is not scheduled because he does not have any queued packets, it is not considered to be an outage event. As in section V-A, we assume that the data sent to a RS is immediately forwarded to the relayed user, within each frame, and that the time is optimally divided within each subchannel between the BS–RS and RS–MS links. Since the outage is based on the number of users who have not received a rate higher than a reference rate R, we define our algorithm as follows: in each subchannel, all users are ordered according to their priority metric on all the access points, and the subchannel is allocated to the best user. The priority metric is ϕ defined in Eq. (1), but here the average user rate $\bar{\beta}_k(t)$ is updated after every subchannel allocation, in order to obtain a finer allocation, namely

$$\bar{\beta}_k(t, n_c) = \frac{p-1}{p} \times \bar{\beta}_k(t-1) + \frac{1}{p} \times \sum_{n=1}^{n_c} c_{k,n} r_{k,n}(t) \tag{11}$$

with n_c the currently served subchannel. In the single RS case, the user on the link with the best ϕ is scheduled in each subchannel. This algorithm is denoted *Optimized Outage* algorithm.

VI. CSI Reduction with Proposed Algorithms

Usually, a major concern for RS–aided allocation is the increased amount of CSI. An optimal algorithm performed at the BS requires:

1) the CSI for BS–MS link of all K users in the cell per subchannel per frame
2) the CSI for RS–MS link of all K users in the cell per subchannel per frame.

This results in a tremendous amount of overhead, especially as the number of RS increases. In the single RS case, for *FTD* algorithm, the required feedback is composed of:

1) the CSI of the K_D direct users at the BS, per subchannel per frame, where $K_D \leq K$
2) the CSI of the K_R relayed users at the RS, per subchannel per frame, where $K_R \leq K$

3) user IDs of users in the list U_{Req}, sent from RS to BS. Since $K_D + K_R = K$, the amount of information of $1) + 2)$ for *FTD* algorithm is equivalent to that of $1)$ for the optimal algorithm. It is also equal to the amount of CSI required for the Max CSI algorithm without relay. Since the number of users in U_{Req} is usually small, $3)$ will be reasonably small. Thus, the CSI information required by *FTD* algorithm is much lower than the one required for the optimal algorithm and slightly higher than for Max CSI. Thanks to the requests by the relay, the feedback can be minimized. For the *ATD* algorithm, in addition to the feedback of *FTD* algorithm, the CSI of the allocated relayed users is required at the BS, as it has to compute the throughput of the RS–subframe for time adaptation. But it is still much lower than the information needed for the optimal algorithm, since $K_{R,alloc} \leq K_R \leq K$ where $K_{R,alloc}$ is the number of users allocated in the RS–subframe among the relayed users. In the multiple RS case, the feedback for *MRPA* and *MRAA* algorithms is also equivalent to the *ATD* algorithm, but from each RS. In either case, the CSI required by the proposed algorithms is much reduced compared to the optimal case, since each RS makes its own allocation and forwards only the useful information to the BS.

VII. Numerical Results

A single cell with a BS and one/multiple RS is considered at first, followed by multi–cell environment. Simulations are made over 150000 sets of channel realizations, where user locations are kept constant for a fixed number of channel realizations, then regenerated. Such a large number of realizations is required for the calculation of the system outage, for which a sample of the number of users in outage is taken every 100 frames (see explanations below). The cell has a 1000 m radius and the relays are placed 800 m away from the BS. The path loss model proposed in [20] is used for the three links: BS–MS, BS–RS and RS–MS. Log–normal shadowing with 0 dB mean and 8 dB standard deviation is assumed for the BS–MS links, and with 6 dB standard deviation for the RS–MS links, as the RS–MS distances are in general smaller than the BS–MS distances, thus having a higher probability to be in LOS [17]. Relays are assumed to be deployed in such a way that the effect of shadowing on the BS–RS links is negligible. The multipath fading channel model in [21] is used. The BS power and RS power are fixed to 20 Watts and 5 Watts [17], respectively. Given the subframe, there is equal power distribution in each subcarrier. There are 48 subcarriers and 12 subchannels, each composed of 4 contiguous subcarriers. The frame duration is fixed to 12 ms. Packets arrive at the BS queue following a Poisson process.

A. User Generation in the cell

One of the major motivations for introducing relays is to increase the coverage of a cell, e. g., to support more users located in the outskirts of the cell. While the per–user throughput for the users located in the cell edge may increase when there are relays, it is not clear what the impact will be on the overall throughput performance, as more radio resources are used to support the additional relayed link. Therefore, it is reasonable to think that the performance of the algorithms will depend on the user distribution. Thus, two cases of user distribution have been considered, the uniform distribution and the edge distribution, where users are located towards the cell edge. That is, the situation where users are concentrated towards the cell edge is the most challenging for supporting the communication, but also where the benefit of relays may be most visible. Depending on the single or multiple relay case, the user distribution is modeled as follows.

1. Case of single relay: in this case, we consider the simplified system model depicted in Fig. 6. In the case with single RS, there is not a justifiable 2–D scenario in which the single RS will improve the coverage for the whole 2–D cell. In order to demonstrate the coverage improvement with a single relay, we have opted to generate users along the line that connects the BS and the RS. On the other hand, if users are generated over the whole cell area, then we should consider multiple RS. Therefore, in the case of multiple RS, we will generate the users over the 2–D cell area. In the first distribution, users are uniformly generated from $x = 0$ to the cell edge $x = R_c$, where x denotes the distance from the BS. For the second distribution, a simplified model for edge distribution is adopted. That is, users are uniformly generated along the segment far from the BS, namely from $x = 0.4 \times R_c$ to $x = R_c$, while no users are generated in the segment close to the BS, from $x = 0$ to $x = 0.4 \times R_c$. This model is not realistic since usually, the user density should continuously vary over the axis. However, this model provides a simple approximation when users are distributed towards cell edge, and can still give an idea of the impact of the distribution on the performance metrics.

2. Case of multiple relays: in this case, a non–uniform user distribution model derived from [22], *clustering to the edge of the cell*, is applied. The cell is divided into square bins as shown in Fig. 7, which are each attributed a certain probability to be selected, P_m, for $m \in [1..16]$. To generate more users towards the edge compared to the center, the bin probability increases as the distance from the bin center to the BS increases. The bins sharing the same center to BS distance are characterized by the same bin probability. In this case, 3 regions can be defined where each region corresponds to bins of equal probability:

- the central area C regrouping bins 6, 7, 10 and 11,
- the closer edge area E_1 regrouping bins 2, 3, 5, 8, 9, 12, 14 and 15,
- the further edge area E_2 regrouping bins 1, 4, 13 and 16.

Users are allowed to be generated anywhere in the whole square grouping the 16 bins, including outside the hexagonal cell. The user generation is carried out similarly as described in [22]. First, a bin is selected by sampling the CDF of the bin probabilities with a uniformly distributed random number between 0 and 1. Then, a user is generated randomly within the selected bin area.

Once users are generated either uniformly or towards the edge, each user is attached either to the BS or one of the RS, with the path selection method described in section II-B. After that, time/subchannel allocation is performed. The path selection is renewed each time the average user SNR changes.

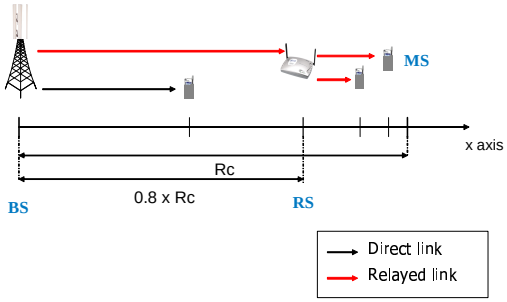

Fig. 6. System Model for the case with Single Relay

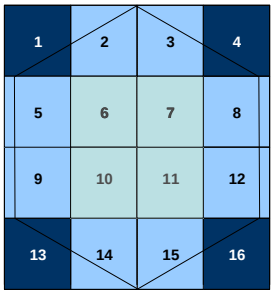

Fig. 7. Hexagonal cell divided into bins, for generating the user distribution towards cell edge, case of multiple relays

B. Performance Metrics

The system performance is characterized by means of two metrics: the goodput and the system outage. The goodput γ in [b/s/Hz], where the overhead for CSI feedback is included, is defined as

$$\gamma = \tau \times \frac{n_{data}}{n_{data} + n_{OH}}, \tag{12}$$

where τ is the throughput, n_{data} the number of OFDM symbols in the frame carrying data and n_{OH} the number of symbols carrying the CSI, assuming QPSK modulation.

The system outage is defined as the probability that the allocated user rates $\overline{r_k}$ are lower than a reference rate R, where $\overline{r_k}$ is averaged over $p = 100$ frames. The system outage probability P_{out} is expressed as

$$P_{out} = \frac{\sum_{s=1}^{S} K_s}{K \times S}, \tag{13}$$

where K_s denotes the number of users in outage for the sample s and S is the total number of samples, K_s =Card$\{k, \overline{r_k} < R\}_s$, where Card denotes the number of elements in the set. If P is the total number of frames, $S = P/p$ since the number of users in outage is taken every p frames.

C. Single RS Case

The FTD and ATD algorithms are compared with the reference algorithms defined in section V: the *RS–Max Optimized* algorithm which gives the throughput upper bound,

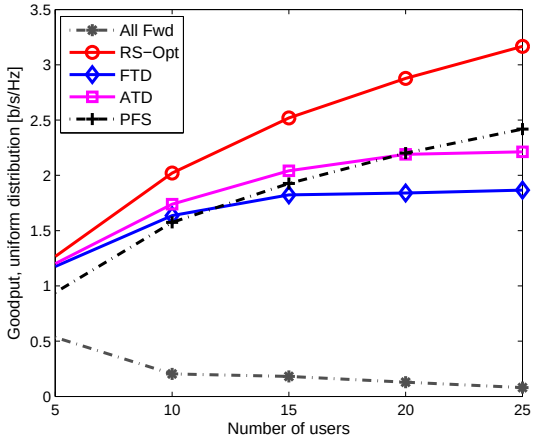

Fig. 8. Cell goodput in [b/s/Hz] for proposed and reference algorithms, with uniform user distribution, case of single relay

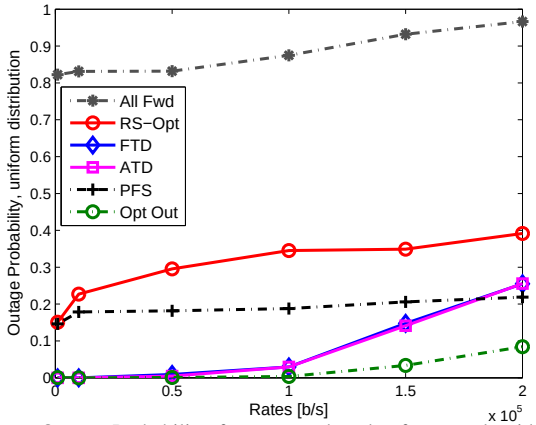

Fig. 9. Outage Probability for proposed and reference algorithms, with uniform user distribution, case of single relay

and *Optimized Outage* for the system outage lower bound. As our algorithms are PFS–based, we also compare them with the *PFS* algorithm without relays, where simply the user with the best ϕ metric is allocated in each subchannel. The achieved goodput is evaluated for 5 to 25 users, and the number of users is fixed to 20 for the outage. When assuming uniform user distribution, Fig. 8 shows that the FTD algorithm achieves a lower throughput than *PFS*, while ATD algorithm has a slightly higher throughput up to 20 users. There is a gap between the goodput of ATD algorithm compared with the one of *RS–Max Optimized* algorithm, which can be explained by the fact that *RS–Max Optimized* algorithm only optimizes the throughput without consideration of PF. The system outage performance of the proposed algorithms is much lower compared to *PFS* and closely follows the outage lower bound given by *Optimized Outage*, except for $R = 200$ [kbps] as shown in Fig. 9. Also, the outage probability of the *RS–Max Optimized* algorithm is unacceptably high.

However, the curve tendency changes when users are distributed towards the cell edge. Fig. 10 shows that both FTD and ATD algorithms drastically improve the cell goodput compared to *PFS* algorithm. The outage performance for FTD and ATD algorithms become also lower than *PFS* for all values

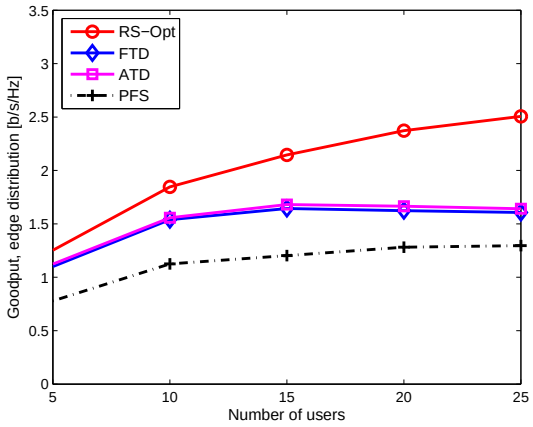

Fig. 10. Cell goodput in [b/s/Hz] for proposed and reference algorithms, with user distribution towards the cell edge, case of single relay

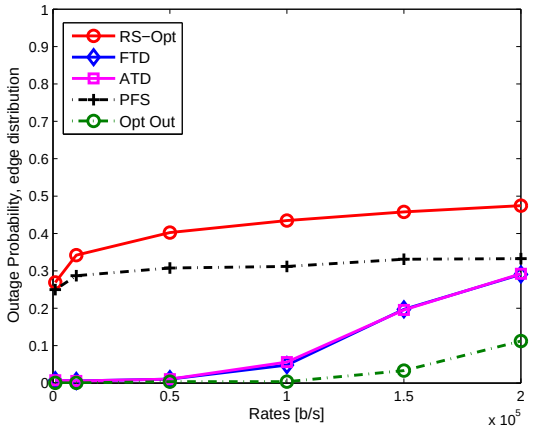

Fig. 11. Outage Probability for proposed and reference algorithms, with user distribution towards the cell edge, case of single relay

of R in Fig. 11. Compared to the uniform distribution, the performance of the *PFS* algorithm becomes very poor for both measures. Note that since *Optimized Outage* is infeasible, the gap of the proposed algorithms with a practical optimal algorithm would be even smaller. Hence, our algorithms can increase the coverage by increasing the number of users satisfying their rate requirement R. The figures show that the utility of the relays is even higher when users are located towards the cell edge than when uniformly distributed. Another interesting point is that, when comparing the performance metrics obtained by uniform distribution and edge distribution, the performance in the latter case drops, for all algorithms including the upper and lower bounds. This is because the direct link quality of all users is lower in average since they are farther from the BS. Thus, the overall throughput and outage deteriorate since less users are attached to the BS with high link quality and more users need to be supported by the RS.

From the figures, it can be observed that with uniform distribution, by adapting the time division of T_{BS} and T_{RS} depending on the queue status at the BS and RS, ATD algorithm significantly improves the throughput while keeping the same outage behavior as the fixed allocation of FTD algorithm. This improvement comes at the price of a higher computational complexity, but which is still much reduced compared to a full search over all the possible time divisions, as discussed in section III-B. On the other hand, ATD and FTD achieve a similar goodput when users are distributed towards cell edge. This is due to the higher utilization of the RS–Subframe with the increased number of relayed users, thereby improving the overall throughput even with a fixed frame. Thus, FTD is well suited for practical use as it achieves a good throughput/outage performance and outperforms *PFS* with lower complexity and required amount of information.

Next, the effectiveness of the relay requesting scheme is evaluated and compared to the *All Forward* algorithm presented in section III-C, for uniform user distribution. Figs. 8 and 9 show that there is a tremendous gain in throughput and in outage with the FTD algorithm, respectively. From Fig. 8, it can be seen that as the number of users grows, more and more users are relayed, so that less and less direct users can be allocated in the BS–subframe, which contributes to the overall downfall of the throughput. This confirms the fact that the BS–subframe is over flooded by the relayed users packets, so that most of the subchannels are allocated to the BS–RS link in detriment of the direct users, resulting in this poor performance. By letting the relay choose and request the relayed users' packets, the subchannel allocation between the BS–RS link and the direct users is optimized, which enables a huge performance gain.

D. Multiple RS Case

With the frequency reuse carried out in *MRPA*, some subchannels are used at the same time, so the opposite relays transmitting in parallel interfere with each other. The effect of this interference is taken into account in the simulations. Namely, if user k attached to RS_1 was scheduled on a subchannel n, then the interference is equal to the signal power of RS_4 to user k, on subchannel n. For the sake of simplicity, the interference is assumed to be an additive Gaussian noise, in which case the SINR of user k on subchannel n denoted $SINR_{k,n}$ can be written

$$SINR_{k,n} = \frac{SNR_{k,n}}{1 + SNR_{k,n} \times \frac{|h_{k,n}^{RS_4}|^2}{|h_{k,n}^{RS_1}|^2}}, \qquad (14)$$

where $SNR_{k,n} = \frac{p_{k,n} \times |h_{k,n}^{RS_1}|^2}{\sigma^2}$ is the SNR of user k on subchannel n. While the subchannel allocation is made based on the SNR values, the SINR values $SINR_{k,n}$ are used to determine the BER values, and thus the achieved throughput.

Results for goodput and outage probability are plotted for varying number of relays, $I = 3, 6, 8, 10, 12$, with $K = 20$ users. For the outage probability, the target rate is fixed to 100 kbps. In these evaluations, the amount of resource used for the preambles and the user mapping information vary with the number of relays. The impact of these control fields is taken into account in the overall throughput. Simply, there is one preamble per relay which occupies one OFDM symbol, $n_{pre} = 1$. In case of frequency reuse by opposite relays, their corresponding preambles also reuse the frequencies. The DL mapping information consists of the ID of the allocated user (or relay for the BS–RS link), for each subchannel.

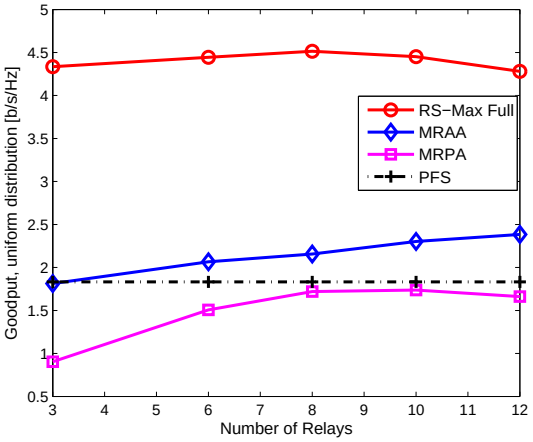

Fig. 12. Cell goodput in [b/s/Hz] for proposed and reference algorithms with varying number of relays, uniform user distribution

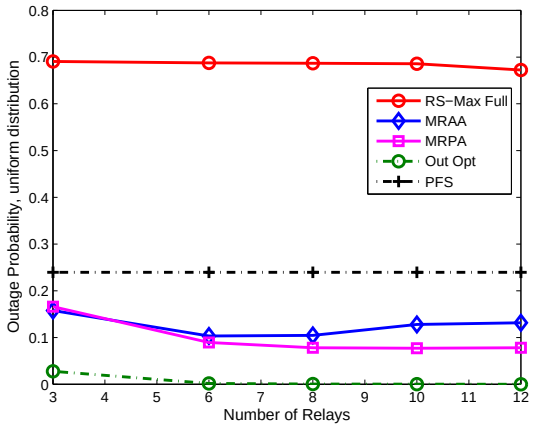

Fig. 13. Outage Probability for proposed and reference algorithms with varying number of relays, uniform user distribution

Denoting n_{MAP} the number of OFDM symbols used for the DL mapping, the goodput is determined as

$$\gamma = \tau \times \frac{n_{data}}{n_{data} + n_{OH} + (N_{RS}+1) \times (n_{pre} + n_{MAP})}, \quad (15)$$

where N_{RS} is the number of non–overlapped, activated RS–Subframes.

Figs. 12 and 13 show the system goodput and outage performance for uniform user distribution, respectively. The goodput and outage of *PFS* algorithm remain constant since this algorithm is independent from the number of relays in the cell. It is observed from Fig. 12 that *RS–Max Full Buffer* algorithm achieves the best goodput when $I = 8$ relays. The goodput decrease at $I = 10, 12$ relays is due to the higher amount of signalling with the increased number of relays.

It is observed that for the proposed algorithms and *Optimized Outage*, the performance of both metrics do not necessarily improve as the number of relays increase. For *Optimized Outage*, the outage at $I = 6$ is already so low that increasing the number of relays does not provide noticeable improvement. *MRPA* achieves the best outage performance for all number of relays and outperforms *PFS*, but its best goodput occurs for $I = 8, 10$ relays. When the number of relays is increased to $I = 12$, the time allocated to each RS_j–subframe is reduced, thereby decreasing the overall throughput while the same outage level is kept. The worst performance of *MRPA* occurs for $I = 3$, which can be explained by the fact that there is no frequency reuse. The improvement in goodput and outage for $I \geq 6$ stresses the benefit of frequency reuse, which gain is larger than the cost of the interference among opposite relay pairs defined in (14). *MRAA* outperforms *PFS* in terms of both goodput and outage, for all number of relays. It can be observed from Fig. 12 that the goodput of *MRAA* increases with the number of relays, even though the amount of signalling increases as shown in (15), which indicates the robustness of the proposed scheme to increasing signalling overhead. At the same time, the outage probability of *MRAA* increases as the number of relays grow for $I \geq 8$, since this algorithm is designed to allocate in each frame, the relays which increase the overall throughput. As the number of relays grow, the number of non–allocated relays increase, causing the outage to rise. Still, the outage performance of *MRAA* at $I = 12$ is considerably lower than the one of *PFS*. Besides, note that *Optimized Outage* is an infeasible scheme so that the difference with the actual optimal scheme is smaller.

When users are distributed towards the edge, the overall performance of all the algorithms is again generally degraded, as observed in Figs. 14 and 15. For all number of relays, while the outage performance of *PFS* is largely degraded, the outage of the proposed algorithms is only slightly lower than with uniform user distribution, pointing out their robustness regarding varying user distributions. Comparing the proposed algorithms with *PFS*, the same conclusions can be made as in the uniform user distribution case. The gain in goodput of *MRAA* compared to *PFS* is even higher than with the uniform distribution, and now *MRPA* also achieves higher a throughput than *PFS* for $I \geq 6$, while achieving the best outage for all I.

The simulations with different number of relays in the cell showed the efficiency of the proposed algorithms, which are robust to varying number of relays and to different user distributions. The *MRPA* scheme achieves the best outage performance, while *MRAA* offers significant goodput increase, at the expense of a slightly higher but reasonable complexity. Thus, *MRAA* achieves the best performance with an excellent trade–off between goodput and outage, for different number of users, relays, and is robust to different user distributions. Under the considered assumptions, the optimal number of relays seem to differ depending on the algorithm. *MRPA* achieves its best performance for $I = 8$, for both user distributions. For *MRAA*, $I = 8$ also achieves the best compromise between goodput and outage, but $I = 12$ offers the best goodput level while considerably reducing the outage probability compared to the conventional algorithm. Interestingly, continuously increasing the number of relays may not be the best solution even when there are more users in the cell edge, since, for example, the goodput of *MRPA* is affected by the increasing amount of signalling overhead as more and more relays are activated and less resource is allocated to each RS_j–subframe. As the number of relays increase, the goodput of *MRAA* improves at the expense of the outage since more relays that decrease the overall throughput are removed.

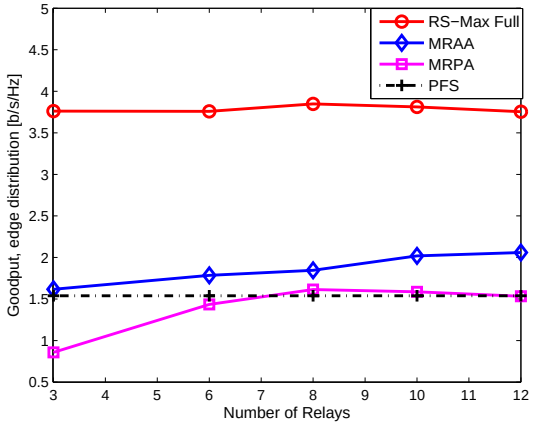

Fig. 14. Cell goodput in [b/s/Hz] for proposed and reference algorithms with varying number of relays, users distributed towards cell edge

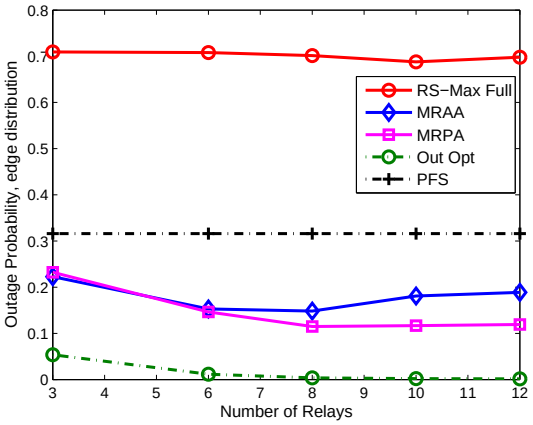

Fig. 15. Outage Probability for proposed and reference algorithms with varying number of relays, users distributed towards cell edge

In addition, the algorithms are evaluated in terms of latency, which is a critical performance measure for real–time services. When using relaying, it becomes extremely challenging to minimize latency, due to the increased number of time frames required to convey the packets to the destination. Here, latency is defined as the actual service time during which a user has to wait before receiving an entire packet composed of 1000 bits. Table I shows the mean and maximum latency values in number of frames for $RS = 8$ for each algorithm. Note that mean and max latencies may not provide the whole information about the latency performance, but they are still important measures of the algorithm performance. As expected, the mean latencies of the proposed algorithms are higher than for *PFS* for both user distributions. When users are distributed towards the cell edge, the latency of *PFS* worsens while the mean latencies of *MRPA* and *MRAA* are not affected as much. Thus, those latency values become very similar. While the maximum latency of *MRAA* is higher than for *PFS* with uniform distribution, the maximum latencies of both proposed algorithms become lower than for *PFS* when users are distributed towards edge. This further confirms the benefit of the proposed algorithms, which provide an excellent trade–off between goodput, outage and latency.

TABLE I

MEAN AND MAXIMUM LATENCY VALUES IN NUMBER OF FRAMES

Mean latency values in number of frames					
	PFS	MRAA	RS–Max Full	MRPA	Out Opt
Uniform	3.0	3.3	6.5	4.0	1.7
Edge	3.6	3.7	7.8	4.2	1.8
Maximum latency values in number of frames					
	PFS	MRAA	RS–Max Full	MRPA	Out Opt
Uniform	3017	4500	6011	2049	503
Edge	5019	4500	10016	4500	502

E. Multi–Cell Environment

Until now, we have only considered allocation within a single–cell, without inter–cell interference. Here, we give an idea about the performance of our algorithms by assuming a multi–cell environment. As solving the whole allocation problem in the presence of interference is very complex, we only assume a simplified interference model as follows. The 6 adjacent cells surrounding the center cell with I relays each are defined as the interfering cells. In each time frame, one relay is chosen randomly in each adjacent cell for $I = 3$, and one pair of opposite relays for $I > 3$. For the sake of simplicity, we assume that the frame is equally divided among the BS–subframe and RS–subframe in each adjacent cell. In the RS–subframe, the randomly chosen relay pair transmits simultaneously for $I > 3$. Thus, the direct users in the center cell are submitted to a constant interference from the 6 surrounding BS, and the relayed users are interfered by the randomly selected relays in each adjacent cell. This can be viewed as a worst case scenario since the adjacent BS and relays are constantly interfering over all subchannels with the direct/relayed users in the center cell, whereas in reality, the interference pattern is intermittent. Figs. 16 and 17 show the goodput and outage probability performance when users are distributed towards cell edge, respectively. Comparing these curves to Figs. 14 and 15 without interference, we can notice that the performance degradation is very low for all the algorithms. Even in this case, our algorithms with multiple relays outperform the PFS algorithm without relays, in terms of outage and throughput. Although a deeper investigation is required for the multi–cell case, these initial results confirm the potential of our proposed algorithms.

VIII. CONCLUSION

In this work, we have investigated the problem of resource allocation for a relay–aided cellular system based on OFDMA. Algorithms for resource allocation have been proposed, for single RS and multiple RS cases. The main specificity of these algorithms is that the RS makes its own initial allocation to minimize the outage, and thereafter the BS optimizes the final allocation in order to improve the overall throughput. This allows a tremendous decrease of the required signalling. In the single RS case, two algorithms have been designed: one operating with low complexity by keeping a fixed time division and the second, with a higher complexity as the time is adapted. In the multiple RS case, we proposed the concept of RS activation, where the frame structure is adapted depending on the active RS. For different number of relays in the

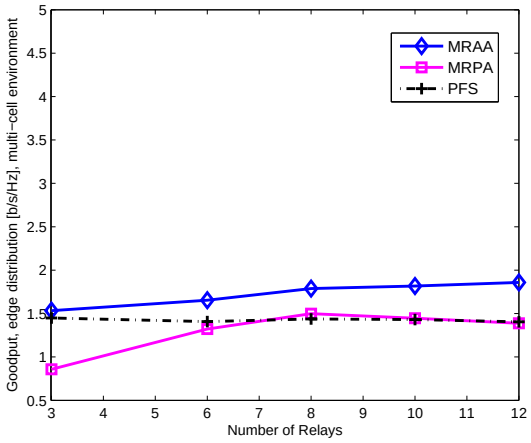

Fig. 16. Cell goodput in [b/s/Hz] in multi–cell environment, users distributed towards cell edge

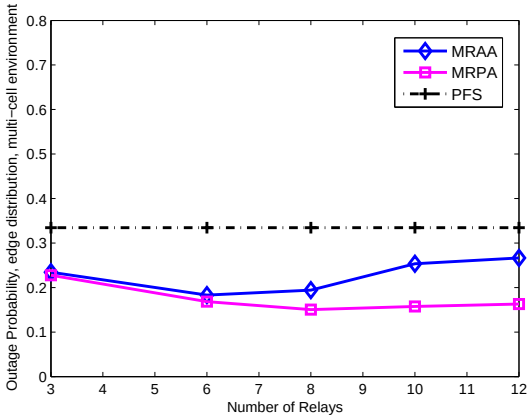

Fig. 17. Outage Probability in multi–cell environment, users distributed towards cell edge

cell, the simulation results show that the proposed algorithms achieve a good trade–off between outage, throughput and latency compared to reference algorithms, while minimizing the computational complexity and required amount of CSI. Depending on the number of relays, the advantages of each algorithm were discussed. Algorithms providing a trade–off between competing performance measures such as throughput and outage are very useful. Moreover, their low level of complexity makes them even practical. Finally, our algorithms provided excellent performance results even in the presence of inter–cell interference. As a future work, these algorithms may be adapted for multi–hop (>2) relaying systems.

ACKNOWLEDGMENT

This work was supported in part by Samsung Electronics, Korea, and by the Grant–in–Aid for JSPS Fellow no. 204205 from the Ministry of Education, Science, Sports, and Culture of Japan.

REFERENCES

[1] R. Pabst, et al., "Relay-Based Deployment Concepts for Wireless and Mobile Broadband Radio," *IEEE Wireless Comm. Mag.*, pp. 80–89, September 2004.

[2] H. Viswanathan and S. Mukherjee, "Performance of Cellular Networks with Relays and Centralized Scheduling," *IEEE Trans. Wireless Comm.*, vol. 4, no. 5, pp. 2318–2328, September 2005.

[3] H. Hu, H. Yanikomeroglu, D. Falconer and S. Periyalwar, "Range Extension without Capacity Penalty in Cellular Networks with Digital Fixed Relays," in *IEEE GLOBECOM*, Dallas, Texas, December 2004.

[4] D. Zheng, J. Zhang and J. Sadowsky, "Hierarchical Multiuser Diversity (HMD) Transmission Scheme," in *IEEE GLOBECOM*, San Francisco, CA, December 2003.

[5] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE ICC*, vol. 1, Seattle, WA, June 1995, pp. 331–335.

[6] J. Jang and K.B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Select. Areas Commun.*, vol. 21, no. 2, pp. 171–178, February 2003.

[7] C. Y. Wong, R. S. Cheng, K.B. Letaief and R.D. Murch, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, October 1999.

[8] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *IEEE GLOBECOM*, San Francisco, CA, December 2000.

[9] P. Viswanath, D.N.C Tse and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Info. Theo.*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[10] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Letters*, vol. 9, no. 3, pp. 210–212, March 2005.

[11] M. Kaneko, P. Popovski and J. Dahl, "Proportional Fairness in Multi–Carrier System: Upper Bound and Approximation Algorithms," *IEEE Comm. Letters*, vol. 10, no. 6, pp. 462–464, June 2006.

[12] Qualcomm, "1× EV: 1× Evolution IS-856 TIA/EIA Standard Airlink Overview," *Revision 7.2*, November 2001.

[13] M. Kaneko and P. Popovski, "Radio Resource Allocation Algorithm for Relay–aided Cellular OFDMA System," in *IEEE ICC*, Glasgow, Scotland, June 2007.

[14] ——, "Adaptive Resource Allocation in Cellular OFDMA System with Multiple Relay Stations," in *IEEE VTC-Spring*, Dublin, Ireland, April 2007.

[15] J. Proakis, *Digital Communications*. McGraw-Hill, 1995.

[16] S.T. Chung and A.J. Goldsmith, "Degrees of Freedom in Adaptive Modulation: A Unified View," *IEEE Trans. Comm.*, vol. 49, no. 9, pp. 1561–1571, September 2001.

[17] I-K. Fu, et al., "Reverse Link Performance of Relay-based Cellular Systems in Manhattan-like Scenario," IEEE 802.16 MMR, $C80216mmr-06\_004r1$, Jan 2006.

[18] M. Asa, et al., "Recommendations for the Scope and Purpose of the Mobile Multihop Relay Task Group," IEEE 802.16 MMR, $C80216mmr-05/032$, Nov 2005.

[19] G. Senarath, et al., "Preliminary Performance Benefit of Single-Hop OFDMA Relay in IEEE 802.16," IEEE 802.16 MMR, $S80216e-05/010r1$, Sept 2005.

[20] S. Ichitsubo, et al., "2 GHz-Band Propagation Loss Prediction in Urban Aeras; Antenna Heights Ranging from Ground to Building Roof," in *IEICE Technical Report, AP 95-15*, May 1996.

[21] S. Yoon, et al., "Orthogonal frequency division multiple access with an aggregated sub–channel structure and statistical channel quality measurements," in *Proc. IEEE VTC*, vol. 2, Los Angeles, CA, September 2004, pp. 1023–1027.

[22] M. Newton and J. Thompson, "Classification and Generation of Non–Uniform User Distributions for Cellular Multi–Hop Networks," in *IEEE ICC*, Istanbul, Turkey, June 2006, pp. 4549–4553.