Fair Resource Allocation Toward Ubiquitous Coverage in OFDMA-Based Cellular Relay Networks With Asymmetric Traffic

Mohamed Salem, *Student Member, IEEE*, Abdulkareem Adinoyi, *Member, IEEE*, Halim Yanikomeroglu, *Member, IEEE*, and David Falconer, *Life Fellow, IEEE*

***Abstract*—Next-generation wireless networks are preoccupied with the provision of very high data rates in a ubiquitous and fair manner throughout the service area. Toward that end, the deployment of fixed relays by the operators has become an accepted network architecture for which orthogonal frequency-division multiple access (OFDMA) is the envisioned air interface, and efficient resource utilization is imperative. In contrast to the current literature, this paper presents a novel throughput-optimal formulation, which performs joint intracell routing and scheduling, in accordance with the emerging OFDMA-based cellular relay networks employing two-hop half-duplex relaying. Low-complexity iterative algorithms are devised to solve the formulated optimization over two consecutive subframes (the base station transmits, followed by the relay stations) using queue-length coupling. We first show that the network capacity, below which the policy is throughput optimal, has been significantly increased, compared with the previously proposed quasi-full-duplex relaying (FDR) scheme, at a slight complexity increase. Hence, throughput fairness and ubiquity have been improved at high traffic loads, aside from the substantial improvement in both queue-awareness and latency. Second, we show that, without empirical priority weights, our efficient implementation of throughput-optimal scheduling achieves a ubiquitous and fair service within each class of users (with symmetric traffic) and across classes of asymmetric traffic in a relative sense on different time scales. Load balancing among only the active relays could still be jointly realized with the resource allocation.**

***Index Terms*—Cellular, fairness, intracell routing, load balancing, orthogonal frequency division multiple-access (OFDMA), radio resource management (RRM), relaying, throughput, ubiquity.**

Manuscript received July 19, 2010; revised November 3, 2010 and January 17, 2011; accepted February 19, 2011. Date of publication March 28, 2011; date of current version June 20, 2011. This work was supported by Samsung Electronics Company, Ltd., Samsung Advanced Institute of Technology, Korea. Patent filings have been made in Korea (application no: P2009-0022132, Mar. 2009) and in the U.S. (application no: 12/567,776, Sept. 2009). International filing is underway. This paper was presented in part at the IEEE Global Telecommunications Conference, Honolulu, HI, November 30–December 4, 2009. This work was done while A. Adinoyi was with Carleton University. The review of this paper was coordinated by Prof. Y.-B. Lin.

M. R. Salem, H. Yanikomeroglu, and D. Falconer are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: mrashad@sce.carleton.ca; halim@sce.carleton.ca; ddf@sce.carleton.ca).

A. Adinoyi is with Swedtel Arabia, Riyadh 11527, Saudi Arabia (e-mail: adinoyi@sce.carleton.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2011.2132746

I. Introduction

RELAYING and orthogonal frequency division multiple access (OFDMA) are among the key technologies for deploying the fourth-generation (4G) and beyond-4G wireless networks that are expected to provide ubiquitous high-data-rate coverage. The synergy in the combined two technologies holds the potential to effectively achieve that objective. These potentials can be exploited from robustness to frequency-selective multipath fading and the inherent multiuser and frequency diversity benefits in OFDMA, in addition to the spatial diversity and routing opportunities in relaying. Moreover, relay deployment provides a cost-efficient way of combating pathloss and expanding the network without incurring the backhaul cost associated with the deployment of additional full-fledged base stations (BSs) [1].

The opportunities in relaying and OFDMA techniques also bring some interesting challenges due to the increased dynamics, degrees of freedom, resource reuse, and complexities incurred in resource allocation and interference management, particularly in networks with large numbers of users and relays [2]. This fact highlights the importance of dynamic and intelligent radio resource management (RRM) schemes with efficient spectrum utilization [2], [3]. As such, the literature on RRM in OFDMA-based cellular relay networks is steadily growing, discussing various schemes in terms of objective (user-centric or network-centric), processing, and feedback (ranging from fully centralized to distributed), as well as scope (considering systems with single cell/single relay to multicellular/multiple relays) [4].

However, the vast majority of schemes proposed so far overlooks some key facts in such environments. First, wireless network traffic is bursty in nature and therefore precludes a one-to-one mapping between channel achievable capacity and user's throughput. Therefore, the applicability of RRM schemes designed to maximize the total achievable capacity, or even to allocate fair shares of this capacity to users, is doubtful in prospective cellular networks. This is because, in reality, such schemes neither deliver a throughput-fair service nor can they exploit the "traffic diversity" (i.e., statistical multiplexing of traffic). Clearly, the inability to maintain fairness defeats service reliability and ubiquity as service becomes channel and location dependent. On the other hand, users pertaining to the same service class could be similarly charged while the service is not evenly distributed [5]. The lack of traffic or queue awareness also prevents such schemes from accounting for previously

relayed data that need to be rescheduled due to a practical automatic repeat request (ARQ) protocol. Second, the radio resource allocation (RRA) problem in such networks is, in principle, a joint routing and scheduling problem rather than just scheduling on preset routes [2].

II. Contributions and Prior Work

Devising dynamic traffic- or queue-aware RRM schemes tackling the joint routing and scheduling problem constitutes therefore a worthwhile yet challenging research opportunity. In [6], Tassiulas *et al.* laid a foundational theory on throughput-optimal scheduling in wireless multihop mesh networks incorporating queue awareness into the scheduling policy, which dynamically allocates resources to multicommodity flows. They showed that maximizing the sum of a queue length-based drift metric over all node pairs is the maximum throughput policy, which stabilizes all network queues under the largest set of mean exogenous arrival rates for which the network queues can be stabilized. Nevertheless, the authors stressed that devising efficient algorithms to solve the optimization problem, given the constraint set imposed by the system model of each particular application, is important for implementation.

Several works have adopted throughput-optimal scheduling, thereafter proposing scheduling policies for adhoc networks, non-OFDMA, or conventional (nonrelaying) cellular networks with different optimization formulations. For instance, in [7] and [8], conventional cellular space-division multiple access/time-division multiple access and OFDMA networks are considered, respectively, thus eliminating the joint routing and scheduling aspect of such policies and limiting the queue-stabilizing opportunities to the resource allocation at the BS.

While fairness is crucial to realize the desired service ubiquity and reliability in cellular networks, it should be noted though that throughput-optimal policies are not fairness oriented in principle, as they aim at stabilizing all user queues under any heterogeneous traffic flows within the system's capacity region. Therefore, in [9], a congestion control mechanism is proposed for a conventional cellular network to introduce user fairness through traffic policing if the arrival rates at the BS are elastic (adaptive).

In [10], Neely *et al.* proposed a centralized dynamic routing and power control (DRPC) policy in a single-carrier ad hoc network with multicommodity flows, rate adaptation, and node power budgets. In each time slot, the DRPC policy solves a *one-shot* optimization to allocate power to a set of links carrying the chosen commodities such that the sum metric is maximized. The authors did not, however, suggest ways to solve such an optimization under the node power constraints and the cochannel interference governing the achievable rates of these links. Therefore, without considering the power control dimension, a centralized joint routing and scheduling algorithm is proposed in [11] for the downlink (DL) of a single-carrier CDMA *cellular* relay network under *symmetric traffic arrival processes*. The authors suppose that throughput-optimal scheduling is a *fair* policy in a such case. It is assumed, however, that a route to the user terminal (UT) may comprise an indefinite number of hops. The algorithm also incurs high complexity and is not applicable to multicarrier systems.

More importantly, the one-shot optimizations in [10] and similar works such as [7], [11], and [12] have no mechanisms to prevent outstanding queues with relatively high mean arrival rates or with the same mean arrival rates yet experiencing high instantaneous bursts from unnecessarily acquiring most—if not all—of the system resources during the subject time slot. This is because the backlog weights will not be affected by how many resources are allocated until the next time slot. Thus, in such scenarios, resources are wasted while low traffic flows experience high latency until their backlog weights dominate; this indeed implies a limitation on the system's capacity (within which the policy is throughput optimal) due to implementation. An enhanced DRPC policy is suggested in [10], where the priorities of low traffic queues are enforced through some empirical scaling parameters at each node.

In our earlier work [13], we formulate a throughput-optimal policy for an OFDMA-based cellular relay network with symmetric traffic at the BS. The policy performs joint in-cell routing and scheduling using only two-hop relaying and prevents resource waste, in contrast to prior art, through efficient bit-loading constraints or iterative optimization in the low-complexity algorithm. We also demonstrate in [13] the system's performance in both the open-routing mode and the practical constrained-routing mode, compared to a relay-enhanced proportional fair scheduler (PFS). However, despite the significant performance returns of the proposed algorithm in [13], it suffers from a performance-limiting bottleneck as the traffic load increases. In addition, relays in that scheme are assumed to be capable of concurrently transmitting and receiving different data on orthogonal frequency-division multiplexing (OFDM) subchannels. This quasi full-duplex relaying (quasi-FDR) raises a practical concern due to the limitations in hardware technology.

Therefore, in this paper, we present a novel throughput-optimal formulation in accordance with the emerging OFDMA-based cellular relay networks *employing half-duplex relaying* (HDR). This paper mainly *generalizes* our initial formulation and algorithms presented in [14], where we show some preliminary performance results against some nonrelaying schemes under only symmetric traffic. Importantly, this paper studies the system's performance under both *symmetric* and *asymmetric* traffic, provides vital and comprehensive discussions on various aspects of the proposed scheme, addresses its practical implementation, and substantiates it from prior art. As such, the research contributions of this paper can be summarized here.

1) A novel throughput-optimal formulation of the RRA problem in next-generation networks is developed for HDR, which is considered realistic for practical implementations [15].
2) Low-complexity iterative algorithms to solve the formulated optimization problem are devised, where the DL RRA over two time slots is separated using *backlog coupling information*, and the connection to the canonical end-to-end achievable capacity is introduced. Dynamic joint routing and scheduling is thus employed, in contrast to most works, e.g., [16] and [17].
3) We show that the network capacity for which the policy is throughput optimal has been significantly increased, compared with the quasi-FDR scheme in [13] and the prior art therein. We also explain how traffic diversity and

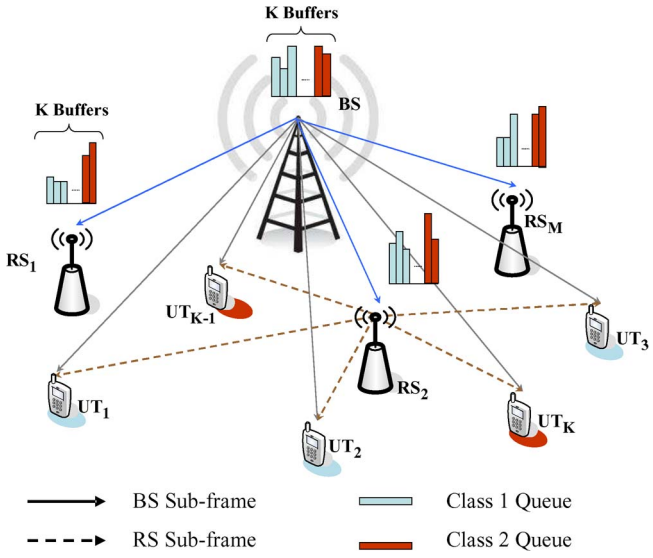

Fig. 1. Representative cell in the multicellular network with asymmetric traffic flows and queue dynamics. The blue and red shades distinguish UTs pertaining to different classes, along with their respective queues.

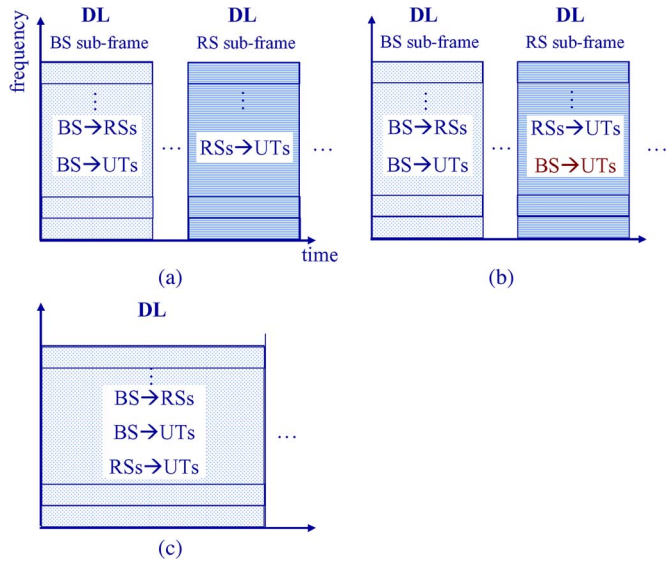

Fig. 2. Generic frame structures for (a) the proposed Variant A, (b) the investigated Variant B, and (c) the quasi-FDR in [13].

queue awareness are better exploited and how the effect of practical ARQ protocols can be taken into account.

4) Our implementation of throughput-optimal scheduling achieves a ubiquitous and fair service within each class of users (with symmetric traffic) and across classes of asymmetric traffic, on the long-term and time-average scales.
5) Load balancing across relay stations (RSs) is achieved jointly with the RRA, as in [13], [18], and [19], yet only among the active RSs; no separate optimization is needed to rearrange the "optimal" solution, in contrast to [20].

The rest of this paper expounds on the previous bullets and is entailed thereafter by Section VII on the implementation issues and the feedback overhead, including the cost of queue awareness at the BS.

III. System Model and Assumptions

We consider a network-level distributed/cell-level centralized RRA scheme [2], using two-hop half-duplex decode-and-forward relaying in the DL transmission of a multicellular network. The BS in each cell communicates with its K UTs, possibly divided into different traffic classes, directly and/or through the assistance of M fixed RSs, which do not exchange traffic with each other. Based on the routing strategy, any UT may simultaneously communicate with multiple (parallel) nodes, and therefore, the BS and each of the M RSs have K separate user buffers. Fig. 1 shows a snapshot of these buffers at different cell nodes where the queue lengths are represented by either blue or red bars, indicating, for instance, two different inelastic traffic classes, i.e., $\mathcal{K}_1$ and $\mathcal{K}_2$ such that $\mathcal{K}_1 \cup \mathcal{K}_2 = \mathcal{K}$. This is a typical cellular setup where the traffic of a set of users pertaining to a certain class is generated as independent and identically distributed, following some distribution. The figure also depicts the generic operation of the joint routing and scheduling in two consecutive DL subframes, i.e., the BS subframe followed by the RS subframe. Aggressive resource reuse is adopted so that the same spectrum is available in each cell.[1] The bandwidth is divided into N subchannels. Each subchannel is a set of adjacent OFDM data subcarriers across which the channel fading is flat. The DL frame structure of our proposed Variant-A scheme is shown in Fig. 2(a), and for the sake of illustration and completeness, Fig. 2(b) shows another possible protocol that defines the Variant-B scheme. In any case, the coherence time of the multipath fading channel is assumed to be greater than the DL frame duration.

In the BS subframe (common to both Variant A and Variant B), only the BS transmits to the selected UTs and RSs. In the proposed Variant A, only RSs transmit to the selected UTs during the RS subframe, whereas in Variant B, while the RSs transmit, the BS directly transmits to some UTs that could be different from those of the first subframe. The subframe times may not necessarily be of equal length, and the RRA formulation takes that into account; the subframe division however could be another optimization dimension that is outside the scope of this paper. Note that, according to the 802.16m frame structure in the TDD mode, for instance, the BS subframe is termed "DL Relay Zone" and is followed first by a UL frame and then by the "DL Access Zone," which resembles the RS subframe [21].[2]

Adaptive modulation and coding (AMC) is employed. Therefore, on each subchannel, the achievable transmit rate at a target bit error rate is a function of the subchannel bandwidth and the signal-to-interference-plus-noise ratio (SINR) at that receiving node. Since the scheme is network-level distributed and has no intercell coordination, fixed power allocation per subchannel is considered for BSs and RSs. In terms of link adaptation, it is also known that power adaptation yields marginal returns when

[1] Without loss of generality, this cell could resemble an LTE-Advanced "cell" served by one of the three-directional beams of an eNB.

[2] Although 3GPP's Release 10 for LTE-A has yet to be finalized, a similar scenario has been discussed in a number of recent technical reports, e.g., R1-091412 and R1-083191.

used in conjunction with AMC, e.g., [22]. The link achievable rate can be calculated using [23]

$$R_{i,j,n} = W \log_2 \left(1 + \frac{-1.5\alpha_{i,j,n}}{\ln(5P_e)}\right) \quad (1)$$

where $\alpha_{i,j,n}$ is the received SINR from source i at destination j on subchannel n, considering the intercell interference (ICI) observed in the previous transmission; we explain the robustness of RRA schemes to the ICI uncertainty in [2]. P_e and W are the target bit error rate and the OFDM subchannel bandwidth, respectively. As an alternative to (1), either the Shannon capacity formula (possibly with some practical SINR gap or penalty) or a discrete AMC lookup table can be used. In the latter case, discrete AMC-level indices, rather than the exact SINR value, can be fed back to alleviate the feedback overhead.

IV. Mathematical Formulation of the Radio Resource Allocation

To achieve a ubiquitous and reliable service in such systems, the RRA scheme has to dynamically route and allocate appropriate resources to each admitted user's traffic flow, regardless of the UT's location, instantaneous traffic bursts, and short- and long-term channel conditions. In other words, the scheme has to achieve throughput fairness within each class of symmetric traffic flows and, more importantly, achieve relative fairness across these asymmetric classes such that light traffic flows are not deprived resources due to the heavy traffic flows. These notions of fairness and ubiquity of throughput-optimal scheduling are quite uncommon in the literature due to the absence of a cellular system model and/or the lack of an efficient implementation that reveals such behavior. It is also worth emphasizing that such fairness notion does not contradict with our intuition of the throughput-fairness tradeoff commonly observed in the literature that considers systems with continuous backlogs or full buffers. Therein, the user with the highest achievable rates would always achieve maximum resource utilization and, thus, maximum throughput, if assigned the whole resources on the expense of fairness.

Since, in principle, throughput-optimal policies dynamically perform joint routing and scheduling of traffic without knowledge of the channel and traffic statistics, a maximization of the sum of the drift metric with proper constraints on frame-by-frame basis achieves our throughput and fairness objectives and exploits the degrees of freedom in multiuser, spatial, and traffic diversities as in the quasi-FDR scheme in [13]. However, unlike mesh networks with multicommodity flows, all user traffic flows have to originate from only one node, which is the BS in the cellular model. This indeed means that, at high traffic loads, the first-hop links will create a bottleneck in the quasi-FDR scheme as the resources per DL frame have to be shared with the second-hop links, which forward the previously stored data at the RSs to the UTs [see Fig. 2(c)]. Therefore, to improve the system's capacity and resource utilization and to reduce the minimum delay of relayed packets, the policy has to grant the BS sole access to the resources during the first portion of the DL frame, i.e., the BS subframe in both Fig. 2(a) and (b).

The policy uses the queue lengths at the RSs to form the differential backlog information that decides which user traffic to be routed from the BS (node 0) through RS_m, whereas the achievable rate $R_{0,m,n}$ on that BS–RS_m feeder link determines how many data units will be forwarded to the corresponding user's buffer at RS_m if subchannel n is acquired. We will elaborate in Section V on the dynamics and the learning ability of this routing strategy to avoid RSs with poor access links to the destined UTs, as applicable to cellular systems with only two-hop relaying. What is important to notice here is that such joint policies, in general, deliver optimal throughput while operating as a slot-by-slot dynamic control *with no coupling* between the optimizations across time slots, *except for the queue lengths*. Such a fluid flow approach to routing or relay selection is also uncommon in the literature of cellular relay networks, where optimization of the frame usually involves complex maximization of the end-to-end achievable capacity over two consecutive slots (subframes) defined as $C^m_{e2e}(k) = \frac{1}{T_1+T_2}\min\{\sum_{n_1\in\mathcal{N}_{0\to m,k}} R_{0,m,n_1}T_1, \sum_{n_2\in\mathcal{N}_{m\to k}} R_{m,k,n_2}T_2\}$ and thus, cannot be separated into two optimizations, i.e., one per subframe. $\mathcal{N}_{m\to k}$ denotes the set of subchannels assigned for the access of relayed UT_k at RS_m during the RS subframe of duration T_2, whereas $\mathcal{N}_{0\to m,k}$ denotes the set assigned to the feeder link during the BS subframe of duration T_1.

Utilizing the flexibility in the fluid flow offered by the queue-length coupling across subframes, two separate optimization procedures are performed (for the two subframes) before the BS starts transmitting in the first subframe. The BS sends the allocation results for the second subframe to the associated RSs on separate control channels. During the uplink portion of the frame, RSs may feed back the actual status of only the queues affected by changes that are not known to the BS. In fact, we observe that the queue-length coupling information in such case is inclusive to, and more practical than, the canonical form of the end-to-end achievable capacity $C^m_{e2e}(k)$, which does not consider the buffer states at the BS and accounts only for the new user data in that frame. In contrast, the queue length at the RS resulting from the first optimization q^m_k can be employed in our scheme to allocate resources to the user access link such that $\sum_{n_2\in\mathcal{N}_{m\to k}} R_{m,k,n_2}T_2 \le q^m_k$ and, meanwhile, accounts for older data units residing at the RS and that need to be rescheduled due to a practical ARQ or hybrid ARQ (HARQ) protocol.

A. Joint Routing and Scheduling for the BS Subframe

The joint routing and scheduling optimization at the BS for the BS subframe can be formulated for both variants A and B, as a binary integer linear programming (BILP) problem. As we noted earlier, the drift (or loosely "demand") metric of any BS–RS feeder link on subchannel n incorporates the maximum difference between the queues at the BS and those at the RS. In addition, the queue length at a UT is always zero in the DL model, resembling a traffic flow sink. Therefore, the sum-demand maximization problem is formulated as

$$\max_{\boldsymbol{\rho}^{(1)},\boldsymbol{\gamma}^{(1)}} \sum_{n=1}^{N}\sum_{k=1}^{K} \rho^{(1)}_{0,k,n} R_{0,k,n} Q^0_k + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K} \gamma^{(1)}_{0,m,n\ (k)} R_{0,m,n}\left(Q^0_k - Q^m_k\right)^+ \quad (2)$$

subject to the constraints

$$\rho^{(1)}_{0,k,n} \in \{0,1\} \qquad \forall_{k,n} \tag{3}$$

$$\gamma^{(1)}_{0,m,n\ (k)} \in \{0,1\} \qquad \forall_{m,n,k}$$

$$\gamma^{(1)}_{0,m,n\ (k)} = 0 \qquad \forall k \ni \mathcal{K}^m \tag{4}$$

$$\sum_{k=1}^{K} \rho^{(1)}_{0,k,n} + \sum_{m=1}^{M}\sum_{k=1}^{K} \gamma^{(1)}_{0,m,n\ (k)} \leq 1 \qquad \forall n \tag{5}$$

$$\sum_{n=1}^{N} \rho^{(1)}_{0,k,n} R_{0,k,n} T_1 + \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma^{(1)}_{0,m,n\ (k)} R_{0,m,n} T_1 \leq Q^0_k \qquad \forall k. \tag{6}$$

In (2)–(6), the first term of the objective function represents the potential users' access links directly from the BS whereas the second term represents the potential feeder links. Thus, $\rho^{(1)}_{0,k,n}$ denotes the kth user's binary assignment variable to the BS on the nth subchannel, during the BS subframe, whereas $\gamma^{(1)}_{0,m,n\ (k)}$ is the mth relay binary assignment variable ($m = 1, 2, \ldots, M$ indexing the RSs) to the BS node on the nth subchannel carrying the traffic of user k. The queue length of user k at node m, which is expressed in bits, bytes, or packets of fixed length, is denoted by Q^m_k. This queue length could change based on the allocation decisions of the BS subframe as in (7) and will thus be denoted by the intermediate coupling length q^m_k in the RS subframe's formulation. The function $(\cdot)^+$ is defined as $(z)^+ = \max\{0, z\}$. The constraints in (3) set the optimization variables to binary values. The set of all user flows that can be routed through RS_m is denoted by $\mathcal{K}^m$, which is equal to $\mathcal{K}$ in the open-routing mode and contains only a subset of $\mathcal{K}$ in the constrained-routing mode, in which, for instance, the UT provides feedback for only a preset number of the closest RSs (denoted by M_{cnst}).[3] As such, the constraints in (4) prevent forwarding the traffic of user k on the feeder link of RS_m according to the constrained routing set $\mathcal{K}^m$

$$q^m_k = Q^m_k + \sum_{n \in \mathcal{N}_{0\rightarrow m,k}} R_{0,m,n} T_1 \qquad \forall m \neq 0; \quad k \in \mathcal{K}^m. \tag{7}$$

The constraints in (5) ensure that at most one link is active per subchannel during the BS subframe. Unlike the majority of works in the literature, e.g., [7], [10], and [11], the constraints in (6) prevent outstanding queues in this one-shot optimization from unnecessarily acquiring most—if not all—of the system resources and thus enabling throughput fairness within a class of symmetric traffic, as well as across asymmetric classes. Note, however, that these constraints do not guarantee that a traffic flow will be allocated some or any resources at all; it is rather the role of the joint policy to maintain the stability of all the queues in the system through appropriate routing and resource allocation. As such, resource waste is also avoided, and the system's capacity is therefore improved, compared with prior art. In the next section, we describe the formulation of the RRA for the RS subframe.

[3]Other selection criteria of relay sets could be based on path loss rather than distance only [24].

B. Formulation of Variant A for the RS Subframe

We recall that, in the proposed Variant A, the BS does not transmit during the RS subframe, and only user access links are considered during that subframe. Here, the throughput-optimal policy operates on the coupling queue length information q^m_k, which is updated by the allocation decisions of the BS subframe before the actual DL transmission. It is important to note that, by incrementing the queues at the RSs as in (7), the feeder link traffic is accounted for when allocating resources to the RSs for the second subframe transmission. It is infeasible and noncausal, on the contrary, in the quasi-FDR scheme [13] and the earlier literature to account for feeder link traffic during the same DL frame due to the concurrent transmission on feeder links and RS-UT links. Therefore, the proposed implementation of throughput-optimal policies features better queue awareness and is thus in a better position toward efficient resource allocation and handling delay-sensitive traffic.

The optimization formulation of Variant A for the RS subframe can be stated as

$$\max_{\boldsymbol{\rho}^{(2)}} \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K} \rho^{(2)}_{m,k,n} R_{m,k,n} q^m_k \tag{8}$$

subject to the constraints

$$\rho^{(2)}_{m,k,n} \in \{0,1\} \qquad \forall_{m,k,n}$$

$$\sum_{m=1}^{M}\sum_{k=1}^{K} \rho^{(2)}_{m,k,n} \leq 1 \qquad \forall n \tag{9}$$

$$T_2 \sum_{n=1}^{N} \rho^{(2)}_{m,k,n} R_{m,k,n} \leq q^m_k \qquad \forall_{m,k}, m \neq 0. \tag{10}$$

The binary variable $\rho^{(2)}_{m,k,n}$ assigns subchannel n to UT_k at RS_m during the RS subframe of duration T_2, where $m = 1, 2, \ldots, M$. Again, the constraints in (9) ensure that at most one link is active per subchannel during the RS subframe, whereas the constraints in (10) caps the resources allocated to each flow to enable throughput fairness, avoid resource waste, and rather achieve efficient resource utilization.

Relay fairness is another important aspect of RRM in OFDMA-based relay networks, which is different from user fairness because there are no quality-of-service requirements specific to RSs. In fact, relay fairness, as it appears in the literature, aims at distributing the traffic load almost evenly among RSs so that no RS will be overloaded [19]. In [18], relay fairness is assessed based on the power consumption at the RSs so that the network operates without overloading the battery of one or more RS(s). Note that, if the RS's transmit power per subchannel is fixed, maintaining almost even distribution of subchannels among RSs limits the RS's total transmit power and, thus, its power amplifier rating and the consumption of its battery energy for solar/battery-operated relays. This is particularly important in the context of green wireless networks, where even fixed RSs rely on the solar energy. In addition, a balanced traffic load reduces the packet processing delays at the regenerative RSs and thus alleviates a practical challenge in the implementation of relay-enhanced networks. In [20], a separate optimization is performed to balance the load (approximated by the number of subchannels as in [19]) among the RSs by

rearranging the optimal allocation. In contrast, such feature can be jointly attained with the resource allocation in this formulation by imposing the following constraint assuming uniform distribution of UTs with respect to the geographical deployment of RSs[4]

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\rho_{m,k,n}^{(2)} \geq \mu_A \qquad \forall m \in \mathcal{M}_{\text{act}} \tag{11}$$

whereas $\mu_A = \lfloor N/|\mathcal{M}_{\text{act}}| \rfloor$ is the minimum number of subchannels that should be assigned to each of the active RSs to balance the load, and $\mathcal{M}_{\text{act}} = \{m : m \neq 0, \Sigma_k q_k^m \neq 0\}$. Note that, if $\mathcal{K}^m = \emptyset$, then $\Sigma_k q_k^m = 0$.

However, strict load balancing may not be desired for more practical scenarios with arbitrary distribution of UTs with respect to the RSs and with constrained routing employed since it is unlikely that RSs will handle even traffic, particularly under high asymmetry between classes. Therefore, in Section V, we discuss how this feature could be practically realized and integrated into our proposed iterative algorithms.

Before the following DL allocation instant, the new traffic arrivals A_k that occurred within the interval between these two allocation instants are added to the user queues at the BS buffer, whereas during the uplink frame, the RSs may report back their actual queue lengths to account for any variations in q_k^m due to, for instance, some BS-transparent ARQ/HARQ requests, which result in rescheduling some data units upon an erroneous reception or dropping some expired delay-sensitive packets. These dynamics can be expressed as follows assuming accurate achievable rates:

$$Q_k^0 = q_k^0 + Q_{k,ARQ}^0 + A_k \tag{12}$$

$$Q_k^m = q_k^m - T_2 \sum_{n=1}^{N} \rho_{m,k,n}^{(2)} R_{m,k,n} + Q_{k,ARQ}^m, \quad m \neq 0. \tag{13}$$

C. Formulation of Variant B for the RS Subframe

It is clear from the literature that the transmission protocol of our proposed Variant-A scheme is not the only possible HDR protocol. Therefore, it is interesting to observe the impact of some other transmission protocol on the performance of our RRA formulation, particularly if it provides some insight on the performance gap between the quasi-FDR protocol and the current contribution. Thus, in Variant B, the BS is treated as an RS during the RS subframe and thus takes a share of the resources to directly communicate with some selected UTs. As such, the formulation of the Variant-B scheme during the RS subframe is the same as that of Variant A but with node index m ranging from 0 to M in (8)–(10). The queue length dynamics as a result of the DL transmission and before the following allocation instant will still follow (13) for the RSs, whereas

$$Q_k^0 = q_k^0 - T_2 \sum_{n=1}^{N} \rho_{0,k,n}^{(2)} R_{0,k,n} + Q_{k,ARQ}^0 + A_k \tag{14}$$

applies for the queues at the BS.

[4]Across various scenarios with the proposed scheme employed, marginal performance loss is realized when the load-balancing constraints are imposed.

Having stated our novel RRA problem formulation, achieving the aforementioned objectives, as the authors stress in [6], depends on devising efficient and practical algorithms to realize the proposed schemes as applied to the cellular system and the associated set of constraints. Since the computational complexity of the above BILP formulation is $\mathcal{O}((M_{\text{cnst}}K)^N)$, $M_{\text{cnst}} \leq M$, and given the expected numbers of subchannels, UTs, and RSs in a practical cellular network, it is inevitable to devise suboptimal low-complexity iterative algorithms to circumvent the prohibitive complexity levels. Therefore, we propose the following iterative algorithms to solve the formulated optimization and achieve its throughput and fairness objectives with tolerable polynomial complexity.

V. Realization of the Proposed Schemes Through Low-Complexity Iterative Algorithms

The BS subframe allocation procedure is the same for both Variant A and Variant B. The BS has full access to all of the N subchannels, yet the transmission occupies only a portion of the DL frame duration (see Fig. 2). The demand metric of any BS-UT link on subchannel n is given as

$$D_{n,0\to k} = R_{0,k,n} Q_k^0 \tag{15}$$

and the demand metric of any BS–RS link on subchannel n is expressed as

$$D_{n,0\to m} = R_{0,m,n} \max_{k\in\mathcal{K}^m} \left\{ \left(Q_k^0 - Q_k^m\right)^+ \right\}. \tag{16}$$

We denote the destination of the "best" BS link, i.e., with the maximum demand, out of $K + M$ potential links on subchannel n, as $\hat{j_n}$. The algorithm then finds the highest demand across all the unassigned subchannels, and the associated BS link denoted as $\hat{j}$ is then selected. The algorithm runs another iteration after eliminating the assigned subchannel and updating the associated queue(s). The iterative process stops when subchannels are exhausted or the queues at the BS are evacuated. Using this greedy iterative assignment approach, the sum-demand is maximized in compliance with constraints (3)–(5), whereas the efficiency and fairness-enabling constraints (6) are satisfied by updating the affected queue(s) (at the BS and the RSs if applicable), according to the assigned rates. Therefore, the BS queues with high traffic load are given their natural priority and allocated the subchannels with the highest achievable rates until they come to around the same back pressure of the low traffic queues; then, the joint policy uses the remaining subchannels to stabilize all the BS queues. Note that, in the literature on throughput-optimal policies, an admission control mechanism is usually assumed at a higher level to either grant or deny any of these traffic flows service based on the system's capacity [6].

In the following, we present the pseudocodes of the RRA algorithms for the BS subframe and the RS subframe based on Variant A. In these codes, $\mathcal{U}$, $\mathcal{N}$, $\mathcal{K}$, and $\mathcal{M}$ denote the sets of unassigned subchannels, all available subchannels, UTs, and RSs, respectively. Recall that, in Variant A (as partly explained in Fig. 2), the BS does not transmit at all, and only RSs share the resources to transmit to the selected UTs during the RS subframe. Similar to the BS-UT links in the previous algorithm, the algorithm here finds in each iteration the best link from any

relay RS_m, out of the $|\mathcal{K}^m|$ links to UTs, on the unassigned subchannel n; such maximum is denoted by $D_{n,m}$.[5]

Since only one link will be active per subchannel, the algorithm needs to compare $D_{n,m}$ across all RSs for each subchannel. If the load-balancing constraints are not imposed, then the algorithm assigns per iteration subchannel $\hat{n}$ to the best link from $\text{RS}_{\hat{m}}$, i.e., line (10) in the second pseudocode is replaced by

$$(\hat{n}, \hat{m}) = \arg\max_{n,m} D_{n,m}. \tag{17}$$

However, if the load-balancing constraints are imposed, the algorithm solves an optimal one-to-one assignment problem per iteration to maximize the total demand by applying the Hungarian algorithm [25] to the tall $|\mathcal{U}| \times M$ demand matrix $[D_{n,m}]$. After each iteration, user queues are updated based on the assigned rates. The iteration process continues until all the traffic in the RS queues is scheduled or the subchannels are exhausted. Note that our implementation of the Hungarian algorithm excludes in any iteration the columns (RSs) with all zero entries; this occurs when the RS has no further traffic to be scheduled or it did not receive any traffic at all, i.e., $\mathcal{K}^m = \emptyset$. As such, when the combined load, due to high and low traffic flows, is almost uniform across all the RSs, the one-to-one assignment jointly achieves strict load balancing and equal power consumption among RSs. On the other hand, when some RSs are inactive and some are handling higher traffic loads than the others, that same algorithm will maintain, from iteration to another, the even distribution of subchannels among only the active RSs, including the lightly loaded RSs, which eventually turn inactive; then, the balancing continues in the remaining iterations among the heavily loaded RSs, and the process continues. Such flexibility in the proposed algorithm, due to the iterative Hungarian, makes it more suitable for practical scenarios as the load is autonomously balanced in a relative sense without invoking an additional optimization.

Modifying the RS subframe algorithm to employ the HDR protocol of Variant B is done by simply running the node index m from 0 to M, and thus, the dimension of the demand matrix in any iteration becomes $|\mathcal{U}|$-by-$(M+1)$.

A. Dynamic Routing in the Two-Hop Cellular Relay Network

Routing in the context of mesh networks employing throughput-optimal policies is dynamically performed using the maximum differential backlog from node a to node b, i.e., $\max_k\{Q_k^a - Q_k^b\}$, and the route may comprise an indefinite number of hops. This is undesirable and expensive, particularly in cellular networks operating in licensed bands. It is not also realistic to assume knowledge of the CSI between any arbitrary pair of RSs across all subchannels, and it is also unlikely with uniform relay deployment that all RSs have good links to the UT. Therefore, we restrict the dynamic routing to the commonly adopted setup, i.e., two hops at most, and thus, RSs are not allowed to exchange traffic. Hence, the differential backlog terms take the form $\max_{k\in\mathcal{K}^m}\{Q_k^0 - Q_k^m\}$, where $\mathcal{K}^m = \mathcal{K}$ in the hypothetical open-routing mode (any UT may receive from any RS) and $\mathcal{K}^m \subseteq \mathcal{K}$ in the practical constrained-routing mode. (Only the best RSs are considered for a UT.)

[5]Since there is no interdependence between the links at different (n, m) pairs, maximizing over k for each pair (n, m) does not affect the combinatorial problem, i.e., does not change the optimal solution.

Pseudocode for the BS subframe for both variants

---

1. Initialization: $\mathcal{U} = \mathcal{N}$, update $\mathbf{Q^0} = [Q_1^0 \dots Q_K^0]$ by new arrivals $\mathbf{A}$, update affected queues in $\mathbf{Q^m}$ by feedback and ARQ rescheduling $\mathbf{Q^m_{ARQ}}$.
2. **while** $|\mathcal{U}| \neq 0$ and $\mathbf{Q^0} \neq \mathbf{0}$ do
3. **for** each $n \in \mathcal{U}$
4. **for** $m = 1$ to M
5. $D_{n,0\to m} = R_{0,m,n} \max_{k\in\mathcal{K}^m}\{(Q_k^0 - Q_k^m)^+\}$
6. $\kappa^m = \arg\max_{k\in\mathcal{K}^m}\{Q_k^0 - Q_k^m\}$
7. **end for**
8. **for** $k = 1$ to K
9. $D_{n,0\to k} = R_{0,k,n} Q_k^0$
10. **end for**
11. $D_{n,0} = \max_j\{D_{n,0\to j}\}, j \in \mathcal{K} \cup \mathcal{M}$
12. $\hat{j_n} = \arg\max_j\{D_{n,0\to j}\}$
13. **end for**
14. $\hat{n} = \arg\max_n\{D_{n,0}\}, \mathcal{U} = \mathcal{U} - \{\hat{n}\}, \hat{j} = \hat{j}_{\hat{n}}$
15. **if** $\hat{j} \in \mathcal{M}$ then
16. $\hat{k} = \kappa^{\hat{j}}, b = \min\{Q_{\hat{k}}^0, \lfloor R_{0,\hat{j},\hat{n}} T_1 \rfloor\}$
17. $Q_{\hat{k}}^0 = Q_{\hat{k}}^0 - b, Q_{\hat{k}}^{\hat{j}} = Q_{\hat{k}}^{\hat{j}} + b$
18. **else**
19. $\hat{k} = \hat{j}, Q_{\hat{k}}^0 = (Q_{\hat{k}}^0 - \lfloor R_{0,\hat{k},\hat{n}} T_1 \rfloor)^+$
20. **end if**
21. **end while**

Pseudocode for RS subframe for Variant A

---

1. Initialization: $\mathcal{U} = \mathcal{N}$, $\mathbf{q^m} = \mathbf{Q^m} \forall m$.
2. **while** $|\mathcal{U}| \neq 0$ and $\sum \mathbf{q^m} \neq \mathbf{0}$ do
3. **for** each $n \in \mathcal{U}$
4. **for** $m = 1$ to M
5. $D_{n,m} = \max_k\{R_{m,k,n} q_k^m\}$
6. $\kappa_{n,m} = \arg\max_k\{R_{m,k,n} q_k^m\}$
7. **end for**
8. **end for**
9. $\%$ $\mathbf{D} = [D_{n,m}]$ is the demand matrix.
10. $(\hat{\mathbf{n}}, \hat{\mathbf{m}}) \Longleftarrow \mathbf{Hungarian}(\mathbf{D})$ $\%$ Vectors of indices
11. $\mathcal{U} = \mathcal{U} - \{\hat{\mathbf{n}}\}, N_{\text{assigned}} = |\hat{\mathbf{n}}| = |\hat{\mathbf{m}}|$
12. $\%N_{\text{assigned}} \leq \min\{M, |\mathcal{U}|\}$
13. **for** $i = 1$ to N_{assigned}
14. $\hat{n} = \hat{\mathbf{n}}(i), \hat{m} = \hat{\mathbf{m}}(i), \hat{k} = \kappa_{\hat{n},\hat{m}}$
15. $q_{\hat{k}}^{\hat{m}} = (q_{\hat{k}}^{\hat{m}} - \lfloor R_{\hat{m},\hat{k},\hat{n}} T_2 \rfloor)^+$
16. **end for**
17. **end while**

Consequently, in the open-routing mode, initial accumulation of the user's traffic may occur at some RS(s) with poor links to the UT as such traffic will be neither forwarded to the UT nor

absorbed by another RS. However, the maximum differential backlog exploits the presence of the trapped data at these RSs, indicating the quality of the second-hop links, and reduces the likelihood of forwarding the user's data on such feeder links in the following iterations and allocation instants. In [13], under uniform deployment of RSs and with different M_{cnst}, the narrow performance gap between the open- and constrained-routing modes of the quasi-FDR algorithm demonstrates this inherent learning ability of the routing strategy to avoid routes with poor second hops in the open mode. We stress that the improvement due to constrained routing comes along with substantial savings in feedback overhead due to the eliminated links, as discussed in Section VII. This learning ability of the joint strategy is also inherent in the proposed algorithms in this paper. Thus, aside from the fairness and ubiquity aspects across asymmetric traffic flows, this observation on the routing behavior of throughput-optimal policies, as applied to two-hop cellular relay networks, is also quite interesting since the common understanding is that imposing constraints on the routing options might reduce the capacity of the multicommodity mesh network.

B. Computational Complexity

The computational complexity of Variant-A and Variant-B schemes discussed in this paper is found to be polynomial in the time of $\mathcal{O}(N^2(N+M)^2/4M)$, $M \leq N$, and $\mathcal{O}(N^2(N+M+1)^2/4(M+1))$, $M+1 \leq N$, respectively. These complexity estimates come from the fact that the Hungarian algorithm is of $\mathcal{O}(|\mathcal{U}|^3)$. These are the complexity levels incurred in the second subframe. However, the proposed scheme incurs a slight increase in complexity of $\mathcal{O}((N^2/2)(K+M))$ due to the first subframe allocation, compared with the iterative algorithm of the quasi-FDR reference scheme in [13].

VI. Numerical Results

Table I provides the simulation parameters used in the study. Most of the parameters are taken from the Third-Generation Partnership Project (3GPP) long-term evolution (LTE) release 9 (Case 3) [26] or the WiMax Forum [27], whereas the WINNER C2 channel model [28] is used. In these system-level Matlab simulations, we have considered 19 hexagonal cells with three or six relays, with equal angular spacing, in each cell. The total DL frame length is 2 ms with equal subframe durations ($T_1 = T_2$). The UTs in each cell are uniformly distributed over the cell area. Since throughput-optimal policies can be applied, regardless of the traffic and channel distributions, independent Poisson packet arrival processes are assumed at the BS queues. The average arrival rate for a Class-1 UT is $\lambda_1 = \lambda$, and that for a Class-2 UT is $\lambda_2 = 2\lambda$, where λ is 632 packet/s (188 bytes each).

On top of the 4-dB lognormal shadowing, the BS–RS links experience *time-frequency correlated Rician* fading with a Rician factor of 10 dB. All other links are non-line of sight and experience 8.9-dB independent lognormal shadowing with *time-frequency correlated Rayleigh fading*. The path-loss model is $PL = 38.4 + 10\beta \log_{10}(d)$ dB, where $\beta = 2.35$ for BS–RS links and $\beta = 3.50$ for RS–UT and BS–UT links. Each RS employs an omnidirectional antenna to communicate with UTs and a highly directive receive antennas with a horizontal gain pattern given in [26] to communicate with its BS. The user mobility used for the study is 20 km/h; however, the scheme can support mobility as high as 90 km/h, given the frame structure and the resulting channel coherence time. In each drop, user locations and shadowing realizations are maintained constant, for which the subject schemes need to compensate, while the traffic and the channel vary on frame-by-frame basis as time evolves.

TABLE I
Simulation Parameters

Parameter	Value
BS-BS distance	1 Km
BS-RS distance	0.65 $\times$ cell radius
Minimum BS-UT distance	35 m
BS Tx. antenna gain	15 dB
RS Tx. antenna gain	10 dB
RS Rx. antenna θ_{3dB}	20°
UT Rx. antenna gain	0 dB
Shadowing σ for NLOS links	8.9 dB
Shadowing σ, for LOS links (BS-RS)	4 dB
Rician K-factor for BS-RS links	10 dB
Carrier frequency	2.5 GHz
Total bandwidth	20 MHz
UT mobility	0-90 Km/hr
Maximum Doppler spread for BS-RS links	4 Hz
No. of channel taps for BS-RS links	8
No. of channel taps for other links	6
TDD frame length	2 msec
Downlink : Uplink ratio	2:1
OFDM subcarrier bandwidth	10.9375 KHz
OFDM symbol duration	102.86 μsec
Subchannel width	18 subcarriers
CR-QAM target BER	10^{-3}
Noise power density at Rx. nodes	-174 dBm/Hz
BS total Tx. power	46 dBm
RS total Tx. power	37 dBm

A. HDR Scheme Versus the Quasi-FDR Scheme and Prior Art

We first consider the case where all UTs belong to the same class, e.g., Class 1, and the cellular network thus handles all symmetric traffic flows ($\mathcal{K}_1 = \mathcal{K}$). Fig. 3 shows cumulative density function (CDF) plots of the time-average user throughput across all drops with $K = 30$ UTs and $M = 3$ or 6 RSs per cell. The same amount of resources is provided for all schemes, i.e., the same DL frame length, bandwidth, and total transmit powers. The figure shows that the quasi-FDR scheme, even in its open-routing mode, outperforms the channel-aware-only relay-enhanced proportional fair scheme, which is discussed in [13], and shows that, at that loading level, a significant throughput gain is realized with the proposed Variant-A HDR scheme, indicating a bottleneck in the quasi-FDR. The figure also shows that the performance gap increases as the number of RSs increases; this can be attributed to the capability of the HDR scheme, as opposed to the quasi-FDR at that point, to exploit the potential increase in spatial diversity and, thus, in the system's capacity when more RSs are deployed with closer proximity to the UTs and good feeder links.

This is in line with our understanding that the bottleneck results from the BS taking only a share of the resources to directly transmit to some UTs, as well as forwarding the relayed traffic on the feeder links. As will be shown in Fig. 5, this does not limit the performance at light-to-moderate loadings, whereas at higher loading, this share of resources becomes

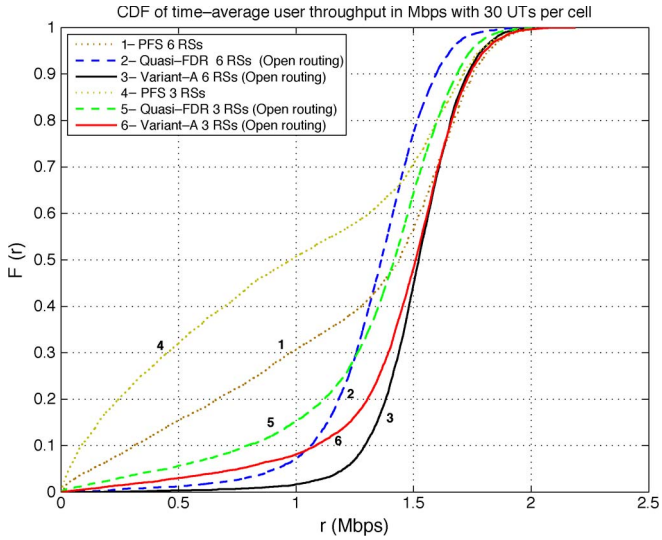

Fig. 3. Time-average throughput comparison of Variant A with the reference schemes at 30 UT/cell.

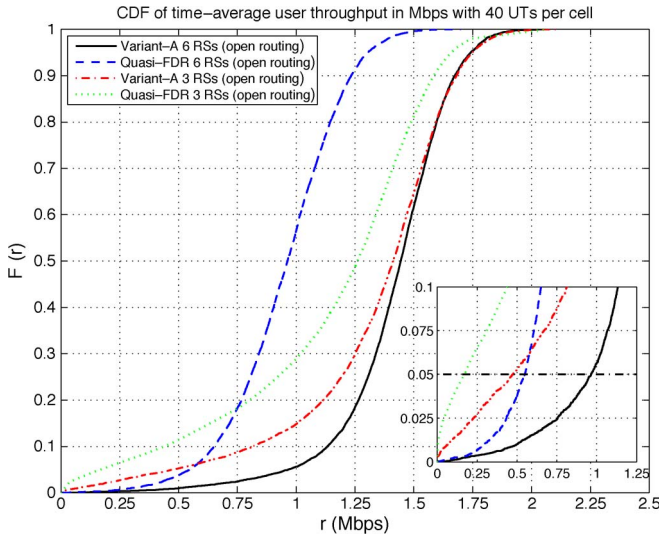

Fig. 4. Total cell average throughput versus the number of UTs per cell for the proposed and reference schemes.

insufficient to serve the traffic. On the other hand, the proposed scheme in this paper grants the BS full access to the whole bandwidth during the BS subframe.

Another informative way of reading these results, according to the LTE evaluation methodology [26], is comparing the cell-edge user throughput attained by the schemes at the fifth percentile. The zoom-in window on the lower tail behavior in Fig. 4 shows that our proposed scheme yields a superior cell-edge performance over the quasi-FDR at a much higher load ($K = 40$). The cdfs of the quasi-FDR scheme with three and six RSs show that reducing the number of RSs relieves the bottleneck at that load to some extent (by increasing the resource share of the BS), thus improving the upper tail. However, the spatial diversity required to enhance the cell-edge throughput is lost, thus affecting the throughput fairness, as shown in Figs. 6 and 7.

The time-average fairness performance of the proposed scheme is presented in Fig. 6 for $K = 40$ using cdf plots of the fairness metric in [29], which can be defined as in (18) with $\beta_i = 1\ \forall i$. Using Jain's index [30], as defined in (19),

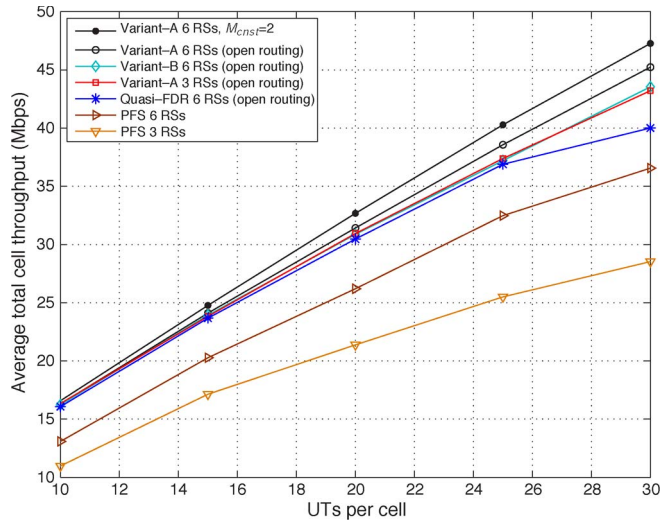

Fig. 5. CDF of time-averaged user throughput with 40 UT/cell and emphasis on the lower tail behavior.

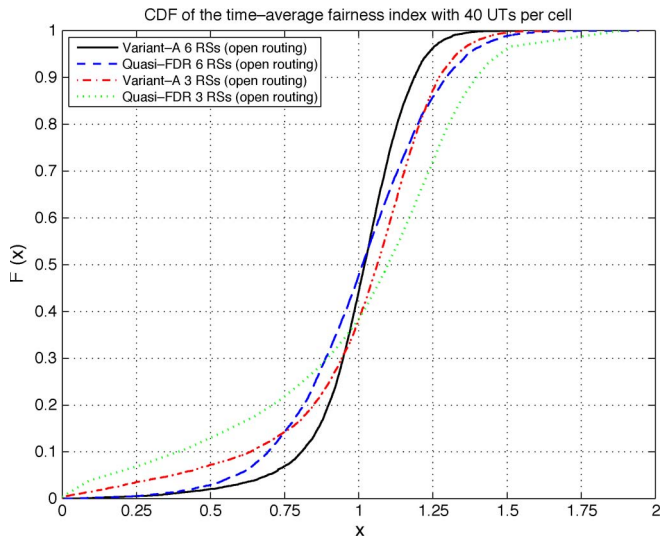

Fig. 6. Time-average throughput fairness for the proposed scheme and the reference quasi-FDR scheme with 40 UT/cell.

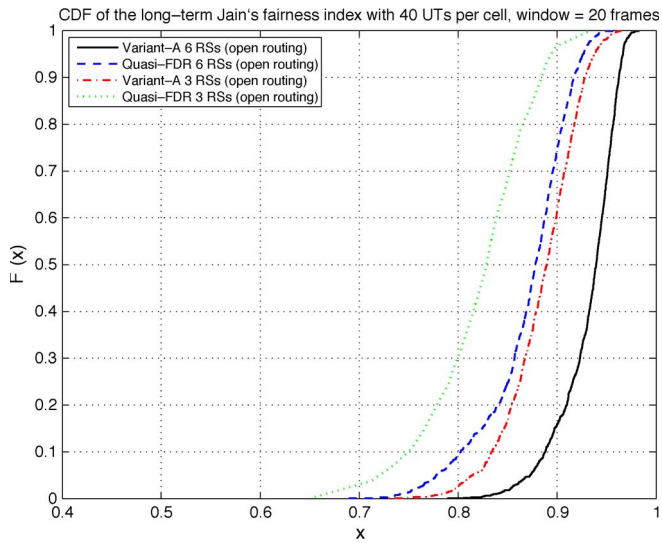

Fig. 7. Long-term throughput fairness for the proposed scheme and the reference quasi-FDR scheme with 40 UT/cell and a time window of 20 frames.

the long-term fairness is demonstrated for $K = 40$ in Fig. 7, where $r_{i,w}$ is the throughput of UT$_i$ during a time window w of 20 frames and $\beta_i = 1\ \forall i$. In these fairness figures, a step function at unity in the cdf plots indicates absolute fairness. Therefore, the closer the curve is to a step function at unity, the fairer the scheme is. It is observed therefore that the proposed scheme achieves the fairest performance, compared with the reference scheme, in the time-average sense and even in the long-term sense. This further underscores the superiority of the HDR scheme in highly loaded networks.

Fig. 5 shows the average total cell throughput as a function of the number of UTs per cell. It is clear that the performance gap between the two variants of the HDR scheme and the quasi-FDR scheme significantly increases as the load increases and becomes insignificant at low-to-moderate loading levels. The impact of the bottleneck in the quasi-FDR scheme can be realized by comparing the slope of these curves at the high loading end, where the proposed scheme outperforms the reference scheme, even with fewer RSs. It is worth mentioning that in Variant B, more resources are devoted to the BS than in Variant A due to the allocation in the RS subframe. However, this results in the scheduling policy becoming less flexible, as the load increases, forwarding the relayed traffic to destined UTs. The performance of Variant B with six RSs is almost the same as that of Variant A with three RSs, yet it is superior to that of the quasi-FDR with six RSs. As discussed earlier in [13], in addition to its substantial feedback savings, constrained routing in two-hop cellular networks enables the joint routing and scheduling policy to achieve better performance in exploiting the deployment geography; this is demonstrated here by the top curve in this figure representing Variant A with six RSs but using $M_{\text{cnst}} = 2$ closest RSs. As such, throughout the rest of our results, the proposed HDR scheme will be represented by Variant A with $M_{\text{cnst}} = 2$. Fig. 4 also shows that the relay-enhanced PFS is significantly inferior to all other schemes; this is due to the lack of traffic or queue awareness and the partitioning of resources and UTs, which is commonly adopted in the literature. A comparison with other nonrelaying schemes can be found in [14], where we also show the latency improvement for relayed packets, compared with the quasi-FDR scheme based on Fig. 2.

B. HDR Variant A With Symmetric and Asymmetric Traffic

We now consider the case where the UTs are equally divided into two groups $\mathcal{K}_1$ and $\mathcal{K}_2$, i.e., Class 1 and Class 2, and the cellular network thus handles asymmetric traffic flows with $K_1 = K_2 = K/2$. The aggregate offered traffic load is represented by the aggregate mean arrival rate $\Lambda = \lambda K/2 + 2\lambda K/2$, which is the aggregate load when all UTs belong to Class 1 such that $K' = 3K/2$. The latter scenario is thus used as a reference scenario, given the same resources and number of RSs.

Fig. 8 shows a scatter plot of user time-averaged throughput as a function of user distance from the BS using Variant A with $K = 20$, $M = 3$ or 6, and $M_{\text{cnst}} = 2$. Each point in the scatter represents the time-averaged throughput for a particular UT within a drop with fixed location and shadowing. The location-based conditional mean is approximated by a 5° polynomial curve fitting as a means of averaging out the effect of shadowing on the joint policy. The figure indicates that uniform average

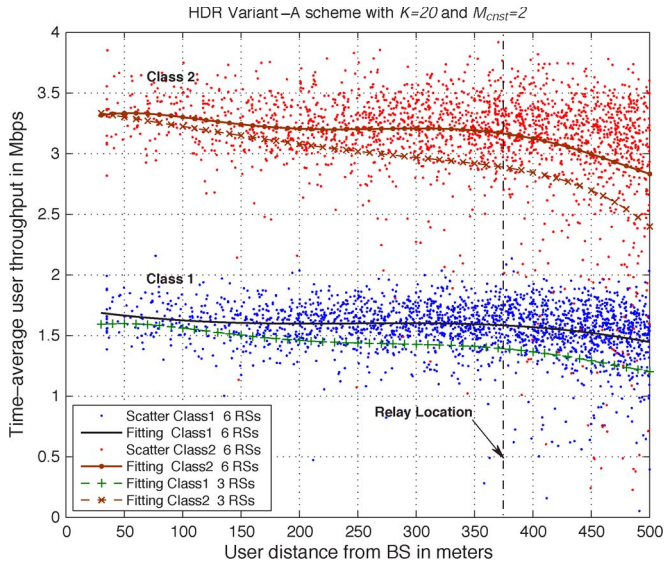

Fig. 8. Time-average user throughput as a function of user location and shadowing with 20 UT/cell with asymmetric traffic and three or six RSs. Other scatters are not shown for figure clarity.

throughput across the cell area is achieved, and thus, a ubiquitous service is attained within each of theses asymmetric traffic classes without imposing priorities on the RRA formulation. This is deduced from the almost flat fittings and the confined spreading of the scatter points. In line with our understanding of the impact of RSs on the capacity and cell-edge performance, the fittings with three RSs show less ubiquity and less cell-edge throughput, compared with the case with six RSs. In both cases, it can be observed that relatively more spreading of the scatter points and less cell-edge throughput are realized for Class-2 UTs, compared with Class-1 UTs.

CDF plots of the scatter points with six and three RSs are shown in Fig. 9(a) and (b), respectively. The lower tail behavior attests to the latter observation on the relative cell-edge performance between Class-1 and Class-2 UTs. To have some insight on the relatively higher spreading (or variance) of Class-2 points, a hypothetical cdf plot is generated by scaling up the time-average throughput realizations of Class-1 UTs by $\beta = 2$. We can generally define the normalizing factor for flow i as $\beta_i = \lambda_i/\lambda_1$. The concordance between the hypothetical cdf and that of Class 2, particularly in terms of variance, and to some extent slope, reveals that the proposed scheme provides almost the same service to the asymmetric traffic flows but in a relative sense, i.e., the realizations of Class-2 service could be roughly approximated by a transformation of the realizations of Class-1 service using the scaling β.

Comparing the cdf of Class-1 UTs in the case of asymmetric load $(K = 20)$ with that of Class 1 in the reference case of symmetric load $(K' = 30)$, it is observed that, with six RSs, the reference curve has an insignificant improvement, mainly at the lower tail. Note that the potential for improvement should be attributed to the increased multiuser/frequency diversity at $K' = 30$. Despite its coexistence with Class-2 traffic, Class-1 traffic receives similar service as that in the all-symmetric case, given the same aggregate load and the same resources. However, with three RSs and, thus, less spatial diversity, the improvement with $K' = 30$ becomes more visible.

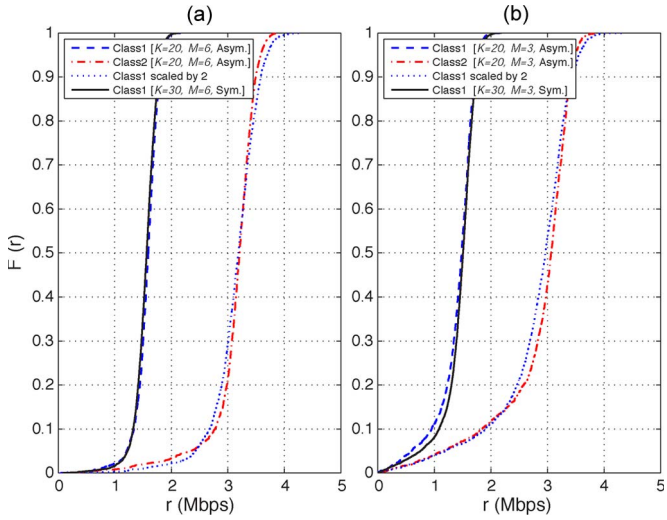

Fig. 9. CDF of time-averaged user throughput with symmetric and asymmetric traffic.

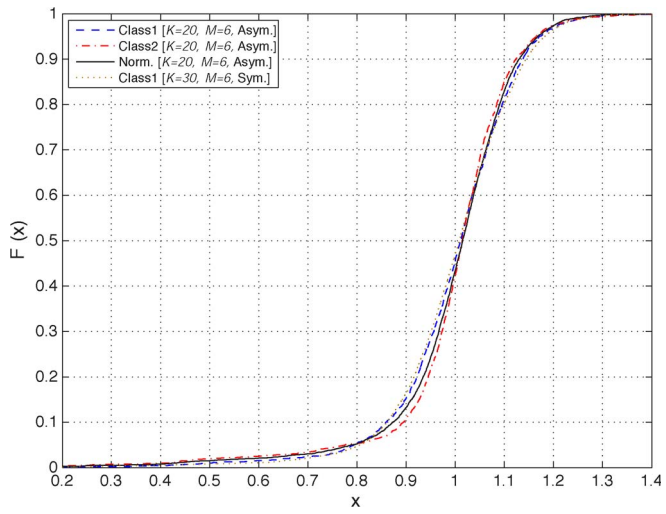

Fig. 10. Time-average absolute and relative fairness with symmetric and asymmetric traffic.

The corresponding time-average fairness performance of the previous cases is shown in Fig. 10 using

$$x_j = \frac{r_j/\beta_j}{\frac{1}{K}\sum_{i=1}^{K} r_i/\beta_i} \tag{18}$$

with $\beta_i = 1\ \forall i$ for absolute fairness within the same class and with the normalized throughput, as in [29] and [31], for relative fairness across the asymmetric classes. Similarly, the long-term fairness is shown in Fig. 11 using Jain's index in

$$x_w = \frac{\left(\sum_{i=1}^{K} r_{i,w}/\beta_i\right)^2}{K\sum_{i=1}^{K}(r_{i,w}/\beta_i)^2} \tag{19}$$

with $\beta_i = 1\ \forall i$ for absolute fairness within the same class and using the normalized throughput as in [32] for relative fairness. In general, all class-based absolute fairness curves are quite close, whereas the relative (normalized) fairness curve also lies in between. The slight improvement in the absolute fairness of Class 2 can be attributed to the less sensitivity of the fairness

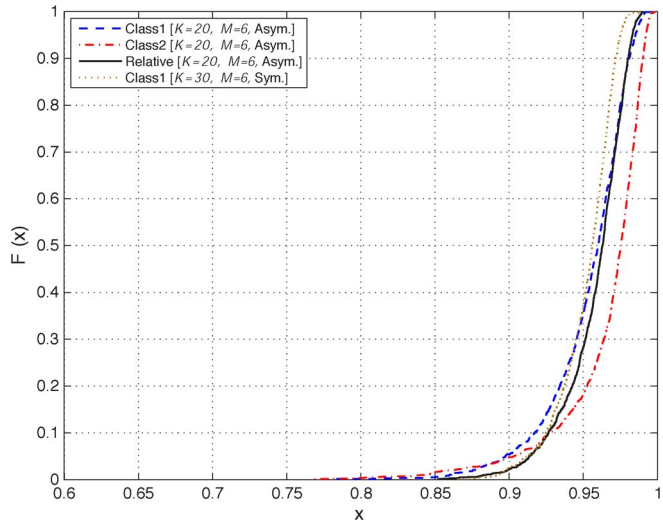

Fig. 11. Long-term absolute and relative fairness with symmetric and asymmetric traffic.

functions at high rate values, along with the slight throughput improvement shown in the cdf of Fig. 9 over the scaled-up throughput of Class 1. This further underscores the superiority of the HDR scheme in highly loaded networks. Once again, the performance of the reference case with $K' = 30$ and six RSs matches that of Class 1 in the asymmetric case.

VII. Implementation Issues and Feedback Overhead

In contrast to traditional cell-level centralized RRM schemes, substantial savings in CSI feedback overhead can be achieved due to the following reasons, which are discussed in more detail in [13]: 1) Implementing the constrained-routing mode reduces the feedback overhead by a factor of $(M_{\text{cnst}} + 1)/M + 1$ since no feedback is required from the UT for the eliminated RS-UT links. 2) Reporting the indices of the achievable AMC levels per link significantly saves in signaling overhead, compared with reporting a wide range of continuous SINRs. 3) Having many UTs per cell and given that a UT can be connected to more than one node, only the "best" fraction (in terms of achievable rates) of the N subchannels needs to be reported per user access link; this reduces the overhead by a factor of N_{CSI}/N[6].

Since queue awareness at the BS is a key element in the proposed RRA algorithms, it is important to investigate whether there is an associated overhead cost, compared with channel-aware only relay-based schemes. Looking at the queue-length dynamics described in (12)–(14), it can be realized that the required queue state results from updating the former state by the new traffic arrivals, the RRA decisions, and finally, the ARQ rescheduling requests, whose overhead is neuter in this paper. As such, we observe that the BS can spontaneously update the queue-length information about its cell nodes (at no cost) or at a minimal cost if system design necessitates, for the following four reasons.

1) Since UTs are the flow sinks of the DL traffic, their queue lengths are set to zero without incurring an overhead cost.

[6]Other results considering only the best 50% of link subchannels show no performance degradation.

On the other hand, the BS is self-aware of its full queue dynamics, including the ARQ requests from the former recipients of its transmissions.

2) In contrast with mesh networks, new traffic arrivals occur only at the BS node in the cellular network, which implies that no exogenous arrivals at RSs or UTs need to be reported to the BS.
3) Since the RRA is cell-level centralized, the BS is aware of the data transmitted from the RSs to the UTs, whereas the relayed data withdrawn from the BS buffers are used by the BS to increment the queue images of the destined RS(s).
4) If a UT generates an ARQ to an RS, the protocol may enable the BS to exploit the broadcast channel to infer the amount of data incrementing back the RS queue and hence update the corresponding image accordingly. If the system design necessitates otherwise or rules out ARQ while channel impairments may cause data losses, then during UL, the RS will need to report the actual change in *only the queues affected by the last DL transmission over its potentially high-speed feeder link*.[7]

It worth noting that the proposed algorithms exploit the finer resource granularity of the HDR frame structure, despite the slight increase in complexity due to the BS subframe. However, at low-to-moderate loading levels, the quasi-FDR scheme achieves the same throughput and fairness performance with the same feedback overhead yet with less computational complexity. Therefore, at such low loading levels, the quasi-FDR is more adequate, provided that advances in technology would have created effective ways to resolve the quasi-FDR implementation challenges.

VIII. Conclusion

Significant throughput fairness and ubiquity can be achieved in a cellular relay network with symmetric inelastic traffic through formulating a throughput-optimal policy that performs joint routing and scheduling on frame-by-frame basis, e.g., the quasi-FDR scheme versus the PFS. We have presented a novel throughput-optimal formulation in accordance with the emerging OFDMA-based cellular relay networks employing HDR. Low-complexity iterative algorithms are devised to solve the formulated optimization over two consecutive subframes using the queue length coupling. Our numerical results have shown that, with a slight complexity increase, compared with the quasi-FDR scheme, the network capacity for which the queues can be stabilized has been significantly increased, and hence, fairness and ubiquity at high traffic loads, aside from the substantial improvement in both queue awareness and latency, are achieved. The results have also shown that, without empirical priority weights, our efficient implementation of throughput-optimal scheduling achieves a ubiquitous and fair service within each class of users (with symmetric traffic) and across classes of asymmetric traffic in a relative sense, on the time-average and long-term time scales. Load balancing among only active relays is jointly realized with the resource allocation.

[7]Furthering the savings in overhead, some quantization of the queue-length process, expressed, for instance, in the number of fixed-size packets or fragments, could be interesting to examine.

Acknowledgment

The authors would like to thank Dr. Y.-D. Kim, Dr. E. Kim, and Dr. Y.-C. Cheong of Samsung Electronics, Samsung Advanced Institute of Technology, Korea, for insightful discussions and invaluable support.

References

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

[2] M. Salem, A. Adinoyi, H. Yanikomeroglu, and D. Falconer, "Opportunities and challenges in OFDMA-based cellular relay networks: A radio resource management perspective," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2496–2510, Jan. 2010.

[3] L. Le and E. Hossain, "Multihop cellular networks: Potential gains, research challenges, and a resource allocation framework," *IEEE Commun. Mag.*, vol. 45, no. 9, pp. 66–73, Sep. 2007.

[4] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y.-D. Kim, E. Kim, and Y.-C. Cheong, "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 3, pp. 422–438, Third Quarter, 2010.

[5] Z. Han and K. J. Liu, *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[6] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[7] M. Kobayashi and G. Caire, "Joint beamforming and scheduling for a multi-antenna downlink with imperfect transmitter channel knowledge," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1468–1477, Sep. 2007.

[8] P. Parag, S. Bhashyam, and R. Aravind, "A subcarrier allocation algorithm for OFDMA using buffer and channel state information," in *Proc. Veh. Technol. Conf.*, Sep. 2005, pp. 622–625.

[9] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1333–1344, Dec. 2007.

[10] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[11] H. Viswanathan and S. Mukherjee, "Performance of cellular networks with relays and centralized scheduling," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2318–2328, Sep. 2005.

[12] M. Neely, E. Modiano, and C. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," in *Proc. IEEE INFOCOM*, New York, Jun. 2002, pp. 1451–1460.

[13] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y-D. Kim, W. Shin, and E. Kim, "Fairness-aware radio resource management in downlink OFDMA cellular relay networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1628–1639, May 2010.

[14] M. Salem, A. Adinoyi, H. Yanikomeroglu, D. Falconer, and Y.-D. Kim, "A fair radio resource allocation scheme for ubiquitous high-data-rate coverage in OFDMA-based cellular relay networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2009, pp. 1–6.

[15] W. Nam, W. Chang, S.-Y. Chung, and Y. Lee, "Transmit optimization for relay-based cellular OFDMA systems," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2007, pp. 5714–5719.

[16] M. Kaneko and P. Popovski, "Radio resource allocation algorithm for relay-aided cellular OFDMA system," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2007, pp. 4831–4836.

[17] Ö. Oyman, "Opportunistic scheduling and spectrum reuse in relay-based cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1074–1085, Mar. 2010.

[18] J. Vicario, A. Bel, A. Morell, and G. Seco-Granados, "Outage probability versus fairness trade-off in opportunistic relay selection with outdated CSI," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, p. 412 837, Jan. 2009.

[19] G. Li and H. Liu, "Resource allocation for OFDMA relay networks with fairness constraints," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 11, pp. 2061–2069, Nov. 2006.

[20] C. Bae and D.-H. Cho, "Fairness-aware adaptive resource allocation scheme in multihop OFDMA systems," *IEEE Commun. Lett.*, vol. 11, no. 2, pp. 134–136, Feb. 2007.
[21] *Syst. Description Document (SDD)*, IEEE 802.16 Broadband Wireless Access Working Group, IEEE Std. 802.16m-09/0034r3, Jul. 2010. [Online]. Available:http://www.ieee802.org/16/tgm/core.html#08_004
[22] Y. Ma, "Proportional fair scheduling for downlink OFDMA," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2007, pp. 4843–4848.
[23] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 884–895, Jun. 1999.
[24] V. Sreng, H. Yanikomeroglu, and D. Falconer, "Relayer selection strategies in cellular networks with peer-to-peer relaying," in *Proc. IEEE Veh. Technol. Conf.*, Oct. 2003, pp. 1949–1953.
[25] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logist. Quart.*, vol. 2, no. 1, pp. 83–97, 1955.
[26] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) System Scenarios*, Apr. 2010. [Online]. Available: http://www.3gpp.org/ftp/specs/html-info/36942.htm
[27] K. Ramadas and R. Jain, "Mobile WiMAX—Part I: A technical overview and performance evaluation," in *Proc. WiMAX Forum*, Sep. 2007, pp. 31–36.
[28] *WINNER II Channel Models*, Mar. 2008. [Online]. Available: http://www.ist-winner.org/WINNER2-Deliverables/D1.1.2v1.1.pdf
[29] *Evaluation Methodology Document (EMD)*, Jan. 2009. [Online]. Available: http://www.ieee802.org/16/tgm/core.html#08_004
[30] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling*. New York: Wiley, 1991.
[31] F. Bokhari, H. Yanikomeroglu, W. K. Wong, and M. Rahman, "Fairness assessment of the adaptive token bank fair queuing scheduling algorithm," in *Proc. IEEE VTC-Fall*, Sep. 2008, pp. 1–5.
[32] A. V. Babu and L. Jacob, "Fairness analysis of IEEE 802.11 multirate wireless LANs," *IEEE Trans. Veh. Technol.*, vol. 56, pt. 2, no. 5, pp. 3073–3088, Sep. 2007.

Mohamed Salem (S'06) received the B.Sc. degree in communications and electronics and the M.Sc. degree from Alexandria University, Alexandria, Egypt, in 2000 and 2006 respectively, and the Ph.D. degree from Carleton University, Ottawa, ON, Canada, in 2010.

He is currently with the Department of Systems and Computer Engineering, Carleton University. He was nominated and hired as a faculty member with the Department of Engineering Mathematics, Alexandria University, where he was promoted to Assistant Lecturer in February 2006. Within this period, he gained wide experience in research and development and collaboration with industrial parties. He has been conducting research in collaboration with Samsung Electronics on advanced radio resource management in next-generation wireless networks. His research interests include stochastic modeling, congestion control, and optimization techniques.

Dr. Salem was granted the 2009/2010 Ontario Graduate Scholarship in Science and Technology.

Abdulkareem Adinoyi (M'07) received the B.Eng. degree from the University of Ilorin, Ilorin, Nigeria, in 1992, the M.S. degree from King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1998, and the Ph.D. degree from Carleton University, Ottawa, ON, Canada, in 2006.

He has worked in the industry (as Engineer, Researcher, and Consultant) and in academia (as Lecturer and Assistant Professor). From January 2004 to December 2006, he worked on the European Union Sixth Framework Integrated Project known as the WINNER. As a Senior Research Associate, he was the project manager of the Samsung-Carleton Collaborative Research Project from 2007 to 2009. He is currently with Swedtel Arabia, Riyadh, Saudi Arabia, as a Consultant for the Saudi Telecommunications Company. His research interests are technology evolution, infrastructure-based multihop and relay networks, cooperative communication techniques and protocols, and radio resource management techniques for broadband wireless networks. He is the holder of three patents (awarded and pending) in radio resource management for relay-based orthogonal frequency-division multiple-access networks.

Halim Yanikomeroglu (M'98) received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 1990 and the M.A.Sc. degree in electrical engineering (now ECE) and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 1992 and 1998, respectively.

He was with the R&D Group of Marconi Kominikasyon A.S., Ankara, from 1993 to 1994. Since 1998, he has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, where he is currently a Full Professor with tenure. He is an Adjunct Professor with Prince Sultan Advanced Technologies Research Institute, King Saud University, Riyadh, Saudi Arabia. He has been a Guest Editor for the *Wiley Journal on Wireless Communications & Mobile Computing*. His research is currently funded by Research In Motion (Canada), Huawei (China), the Communications Research Centre of Canada, and the Natural Sciences and Engineering Research Council of Canada. His research interests include many aspects of the physical, medium access, and networking layers of wireless communications, with special emphasis on multihop/relay/mesh networks and cooperative communications.

Dr. Yanikomeroglu is a registered Professional Engineer in the province of Ontario. He has been involved with the steering committees and technical program committees of numerous international conferences. He has also given 19 tutorials at such conferences. He is a member of the Steering Committee of the IEEE Wireless Communications and Networking Conference (WCNC) and has been involved in the organization of this conference over the years, including serving as the Technical Program Co-Chair of WCNC 2004 and the Technical Program Chair of WCNC 2008. He was the General Co-Chair of the IEEE Vehicular Technology Conference held in Ottawa in September 2010. He was an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2002–2005) and the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS (2002–2003). He was an Officer of IEEE's Technical Committee on Personal Communications (where he was Chair during 2005–2006, Vice-Chair during 2003–2004, and Secretary during 2001–2002) and was also a member of the IEEE Communications Society's Technical Activities Council (2005–2006). He was the recipient of the Carleton University Research Achievement Award in 2009.

David Falconer (LF'06) received the B.A.Sc. degree in engineering physics from the University of Toronto, Toronto, ON, Canada, in 1962, the S.M. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1963 and 1967, respectively, and an honorary doctorate of science degree from the University of Edinburgh, Edinburgh, U.K., in 2009.

After a year as a Postdoctoral Fellow with the Royal Institute of Technology, Stockholm, Sweden, he was with Bell Laboratories from 1967 to 1980 as a Member of Technical Staff and Group Supervisor.

During 1976–1977, he was a Visiting Professor with Linköping University, Linköping, Sweden. Since 1980, he has been with Carleton University, Ottawa, ON, Canada, where he is currently Professor Emeritus and a Distinguished Research Professor with the Department of Systems and Computer Engineering. He was the Director of the Carleton's Broadband Communications and Wireless Systems Centre from 2000 to 2004. He was the Chair of Working Group 4 (New Radio Interfaces, Relay-Based Systems, and Smart Antennas) of the Wireless World Research Forum in 2004 and 2005. His current research interests include beyond-third-generation broadband wireless communications systems.

Dr. Falconer received the 2008 Canadian Award for Telecommunications Research, the 2008 IEEE Technical Committee for Wireless Communications Recognition Award, and the IEEE Canada 2009 Fessenden Award (Telecommunications).