Fair Resource Allocation Towards Ubiquitous Coverage in OFDMA-based Cellular Relay Networks with Asymmetric Traffic

Mohamed Salem, Abdulkareem Adinoyi, Halim Yanikomeroglu, and David Falconer
Department of Systems and Computer Engineering
Carleton University, Ottawa, Canada

Abstract—The next-generation wireless networks are preoccupied with the provision of very high data rates in a ubiquitous and fair manner throughout the service area. Towards that end, the deployment of fixed relays by the operators has become an accepted network architecture for which 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 intra-cell 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 sub-frames (the base station transmits followed by the relay stations) using the queue length coupling. We first show that the network capacity, below which the policy is throughput-optimal, has been significantly increased compared to the previously proposed quasi-FDR scheme, at a slight complexity increase. Hence, throughput fairness and ubiquity have been improved at high traffic loads, besides 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 still could be jointly realized with the resource allocation.

Index Terms—OFDMA, RRM, cellular, relaying, intra-cell routing, throughput, fairness, ubiquity, load balancing.

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 potentials to effectively achieve that objective. These potentials can be exploited from the 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 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, especially 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 burst 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 for maximizing the total achievable capacity, or even allocating 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 charged similarly 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 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-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 allocates resources dynamically 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 (non-relaying) cellular networks with different optimization formulations. For instance, in [7] and [8], conventional cellular SDMA/TDMA and OFDMA networks are respectively considered 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 heterogenous 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 policy (DRPC) in a single-carrier adhoc network with multicmodity 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 co-channel 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 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 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 not applicable to multi-carrier systems.

More importantly, the one-shot optimizations in [10], and the similar works such as [7], [11] and [12], have no mechanisms to prevent outstanding queues -with relatively high mean arrival rates or with 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. That is because the backlog weights will not be affected by how much 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; that 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 priorities of low traffic queues are enforced through some empirical scaling parameters at each node.

In our earlier contribution [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 as 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 transmitting and receiving different data concurrently on orthogonal 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. The paper mainly generalizes our initial formulation and algorithms presented in the conference paper [14] where we show some preliminary performance results against some non-relaying 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 as follows:

- A novel throughput-optimal formulation of the RRA problem in next-generation networks is developed for half-duplex relaying (HDR) which is considered realistic for practical implementations [15].
- Low complexity iterative algorithms to solve the formulated optimization problem are devised where the downlink RRA over two time slots is separated using the 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].
- We show that the network capacity for which the policy is throughput-optimal has been significantly increased as compared to the quasi-FDR scheme in [13] and the prior art therein. We also explain how traffic diversity and queue-awareness are better exploited and how the effect of practical ARQ protocols can be taken into account.
- 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.
- 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 downlink (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 communicate simultaneously with multiple (parallel) nodes, and therefore the BS and each of the $M$ RSs has $K$ separate user buffers. Figure 1 shows a snapshot of these buffers at different cell nodes where queue lengths are represented by either blue or red bars indicating, for instance, two different inelastic traffic classes, i.e., $K_1$ and $K_2$ such that $K_1 \cup K_2 = K$. This is a typical cellular setup where the traffic of a set of users pertaining to a certain class is generated as i.i.d following some distribution. The figure also depicts the generic operation of the joint routing and scheduling in two consecutive DL sub-frames; the BS sub-frame followed by the RS sub-frame. 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 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 sub-frame (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 sub-frame. Whereas in Variant-B, while the RSs transmit, the BS directly transmits to some UTs who could be different from those of the first sub-frame. The sub-frame times may not necessarily be of equal length and the RRA formulation takes that into account; the sub-frame 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 sub-frame is termed ‘DL Relay Zone’ and is followed first by an UL frame then by the ‘DL Access Zone’ which resembles the RS sub-frame [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 inter-cell 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 used in conjunction with AMC, e.g., [22]. The link achievable rate can be calculated using [23]

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

where $\alpha_{i,j,n}$ is the the received SINR from source $i$ at destination $j$ on subchannel $n$ considering the inter-cell 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 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 can be feedback rather than the exact SINR values to alleviate the feedback overhead.

IV. MATHEMATICAL FORMULATION OF THE RRA

In order 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 trade off commonly observed in the literature that

\(^1\)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 not been finalized yet, a similar scenario has been discussed in a number of recent technical reports, e.g., R1-091412 and R1-083191.
consider systems with continuous backlogs or full buffers. Therein, the user with the highest achievable rates would always achieve the maximum resource utilization and thus the maximum throughput, if assigned the whole resources on the expense of fairness.

Since in principle throughput-optimal policies perform joint routing and scheduling of traffic dynamically 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, 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 sub-frame 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 to forward the previously stored data at the RSs to the UTs, frame have to be shared with the second-hop links which maximum difference between the queues at the BS and those of any BS-RS feeder link on subchannel $m$ during the RS sub-frame of duration $T_2$ while $\mathcal{N}_{m \rightarrow k}$ denotes the set assigned to the feeder link during the BS sub-frame of duration $T_1$.

Utilizing the flexibility in the fluid flow offered by the queue length coupling across sub-frames, two separate optimization procedures are performed (for the two sub-frames) before the BS starts transmitting in the first sub-frame. The BS sends the allocation results for the second sub-frame to the associated RSs on separate control channels. During the uplink portion of the frame, RSs may feedback 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, $C_{m \rightarrow k}^f(k)$, 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 BS resulting from the first optimization $q_m^k$ can be employed in our scheme to allocate resources to the user access link such that $\sum_{n \in \mathcal{N}_{m \rightarrow k}} R_{m,k,n} T_2 \leq q_m^k$ and meanwhile accounts for older data units residing at the RS and need to be rescheduled due to a practical ARQ or HARQ protocol.

A. The Joint Routing and Scheduling for the BS Sub-frame

The joint routing and scheduling optimization at the BS for the BS sub-frame 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. Also, 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

$$
\max_{\rho^{(1)}, \gamma^{(1)}} \quad \sum_{n=1}^{N} \sum_{k=1}^{K} \rho_{0,k,n} R_{0,k,n} Q_k^0 + \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_{0,m,n} (k) R_{0,m,n} (Q_k^m - Q_k^m)^+ ,
$$

subject to the constraints

$$
\rho_{0,k,n}^{(1)} \in \{0,1\} \quad \forall k,n, \quad (3)
$$

$$
\gamma_{0,m,n}^{(1)} (k) \in \{0,1\} \quad \forall m,n,k, \quad (4)
$$

$$
\sum_{k=1}^{K} \rho_{0,k,n}^{(1)} + \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_{0,m,n}^{(1)} (k) \leq 1 \quad \forall n, \quad (5)
$$
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, $\rho_{0,k,n}^{(1)}$ denotes the $k^{th}$ user’s binary assignment variable to the BS on the $n^{th}$ subchannel, during the BS sub-frame, while $\gamma_{0,m,n}(k)$ is the $m^{th}$ relay binary assignment variable ($m = 1, 2, \ldots, M$ indexing the RSs) to the BS node on the $n^{th}$ subchannel carrying the traffic of user $k$. The queue length of user $k$ at node $m$, expressed in bits, bytes, or packets of fixed length is denoted by $Q_k^m$. This queue length could change based on the allocation decisions of the BS sub-frame as in (7), and will thus be denoted by the intermediate coupling length $q_k^m$ in the RS sub-frame’s formulation. The function $(z)^+$ 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 $R_{m,n}$ is denoted by $K^m$ which is equal to $K$ in the open routing mode and contains only a subset of $K$ in the constrained routing mode in which, for instance, the UT provides feedback for only a preset number of the closest RSs (denoted by $M_{cnm}$). As such, the constraints in (4) prevents forwarding the traffic of user $k$ on the feeder link of $R_{m,n}$ according to the constrained routing set $K^m$.

$$q_k^m = Q_k^m + \sum_{n\in N_{k-n,m,k}} R_{0,m,n} T_1, \; \forall m \neq 0, \; k \in K^m.$$

The constraints in (5) ensure that at most one link is active per subchannel during the BS sub-frame. 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 as compared to prior art. In the next section we describe the formulation of the RRA for the RS sub-frame.

**B. Formulation of Variant-A for the RS Sub-frame**

We recall that in the proposed Variant-A, the BS does not transmit during the RS sub-frame and only user access links are considered during that sub-frame. Here, the throughput-optimal policy operates on the coupling queue length information $q_k^m$ which is updated by the allocation decisions of the BS sub-frame 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 sub-frame 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 thus in a better position towards efficient resource allocation and handling delay-sensitive traffic.

The optimization formulation of Variant-A for the RS sub-frame can be stated as

$$\max_{\rho_k^{(2)}} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} \rho_{m,k,n}^{(2)} R_{m,k,n} q_k^m,$$

subject to the constraints

$$\rho_{m,k,n}^{(2)} \in \{0, 1\} \; \forall m, k, n,$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \rho_{m,k,n}^{(2)} \leq 1 \; \forall k,$$

and

$$T_2 \sum_{n=1}^{N} \rho_{m,k,n}^{(2)} R_{m,k,n} \leq q_k^m \; \forall k, \; m \neq 0.$$
the geographical deployment of RSs.

\[
\sum_{n=1}^{N} \sum_{k=1}^{K} \rho_{m,k,n}^{(2)} \geq \mu_A \quad \forall m \in M_{\text{act}},
\]

(11)

whereas \(\mu_A = \left\lceil N/|M_{\text{act}}| \right\rceil\) is the minimum number of subchannels that should be assigned to each of the active RSs for balancing the load and \(M_{\text{act}} = \{m : m \neq 0, \sum_k q_k^m \neq 0\}\). Note that if \(K^m = 0\), 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, especially 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 \(A_k\) 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 of some expired delay-sensitive packets. These dynamics can be expressed as follows assuming accurate achievable rates

\[
Q_k^0 = q_k^0 + Q_k^0,_{\text{ARQ}} + A_k,
\]

(12)

\[
Q_k^m = q_k^m - T_2 \sum_{n=1}^{N} \rho_{m,k,n}^{(2)} R_{m,k,n} + Q_k^m,_{\text{ARQ}}, \; m \neq 0.
\]

(13)

C. Formulation of Variant-B for the RS Sub-frame

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, especially if it provides some insight on the performance gap between the quasi-FDR protocol and the current contribution. So, in Variant-B, the BS is treated as a RS during the RS sub-frame, and thus takes a share of the resources to communicate directly with some selected UTs. As such, the formulation of Variant-B scheme during the RS sub-frame is the same as that of Variant-A but with the node index \(m\) ranging from 0 to \(M\) in the equations (8) through (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 (14) applies for the queues at the BS.

\[
Q_k^0 = q_k^0 - T_2 \sum_{n=1}^{N} \rho_{0,k,n}^{(2)} R_{0,k,n} + Q_k^0,_{\text{ARQ}} + A_k.
\]

(14)

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 \(O\left(\left(M_{\text{act}}K\right)^N\right)\), 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 sub-frame 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

\[D_{n,0 \rightarrow k} = R_{0,k,n} q_k^0,\]

(15)

and the demand metric of any BS-RS link on subchannel \(n\) is expressed as

\[D_{n,0 \rightarrow m} = R_{0,m,n} \max_k \{\left(Q_k^0 - q_k^m\right)^+\}.
\]

(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 \(j_n\). The algorithm then finds the highest demand across all the unassigned subchannels and the associated BS link denoted as \(j\) 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 the constraints (3) through (5) while 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 pseudo-codes of the RRA algorithms for the BS sub-frame and the RS sub-frame based on Variant-A. In these codes, \(U, N, K, M\) denote the sets of unassigned subchannels, all available subchannels, UTs, and RSs, respectively. Recall that in Variant-A (as explained partly 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 sub-frame. 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 \(|K^m|\) links to UTs, on the unassigned subchannel \(n\); such maximum is denoted by

\(\text{Max}\)
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 $R_{\hat{n}}$, i.e., line (10) in the second pseudo-code is replaced by
\[
(\hat{n}, \hat{m}) = \arg\max_{n,m} D_{n,m}. \quad (17)
\]

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., $\kappa^m = \emptyset$. 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. Whereas, 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 ones which eventually turn inactive and 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 sub-frame 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 performed dynamically using the maximum differential backlog from node $a$ to node $b$, $\max_k\{Q^k_a - Q^k_b\}$, and the route may comprise an indefinite number of hops. This is undesirable and expensive, especially 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 $\max_k\{Q^k_a - Q^k_b\}$, where $\kappa^m = K$ in the hypothetical open routing mode (any UT may receive from any RS) and $\kappa^m \subseteq K$ in the practical constrained routing mode (only the best RSs are considered for a UT).

Pseudo-code for the BS sub-frame for both variants

1. Initialization: $\mathcal{U} = \mathcal{N}$, update $Q^0 = [Q^0_1 \ldots Q^0_K]$ by new arrivals $A$, update affected queues in $Q^m$ by feedback and ARQ rescheduling $Q^m_{ARQ}$.
2. while $|\mathcal{U}| \neq 0$ and $Q^m \neq 0$ do
   3. for each $n \in \mathcal{U}$
      4. for $m = 1$ to $M$
         5. $D_{n,0 \rightarrow m} = \max_{n,m,n} \max_k \{Q^k_n - Q^m_k\}$
         6. $\kappa^m = \max_{k \in \kappa^m} \{Q^k_n - Q^m_k\}$
      7. end for
   8. for $k = 1$ to $K$
      9. $D_{n,0 \rightarrow k} = \max_j \{D_{n,0 \rightarrow j}, j \in \mathcal{K} \cup \mathcal{M}\}$
     10. $\hat{n} = \max_{n} \{D_{n,0}\}$, $\mathcal{U} = \mathcal{U} - \{\hat{n}\}$
     11. if $\hat{j} \in \mathcal{M}$ then
        12. $\hat{k} = \kappa^j$, $b = \min_{j \in \mathcal{M}} \{Q^k_n, R_{0,j,n}T_i\}$
        13. $Q^0_k = Q^0_k - b$, $Q^j_k = Q^j_k + b$
     14. else
        15. $\hat{k} = \hat{j}$, $Q^0_k = (Q^0_k - [R_{0,k,n}T_i])^+$
     16. end if
   17. end while

Pseudo-code for RS sub-frame for Variant-A

1. Initialization: $\mathcal{U} = \mathcal{N}$, $Q^m = Q^m \forall m$.$\setminus$.
2. while $|\mathcal{U}| \neq 0$ and $\sum Q^m \neq 0$ do
   3. for each $n \in \mathcal{U}$
      4. for $m = 1$ to $M$
         5. $D_{n,m} = \max_k \{R_{m,k,n} q^m_k\}$
         6. $\kappa^m = \max_k \{R_{m,k,n} q^m_k\}$
      7. end for
   8. end for
   9. $\% D = [D_{n,m}]$ is the demand matrix.
   10. $\hat{n}, \hat{m} \leftarrow \text{Hungarian}(D)$ \% Vectors of indices
   11. $U = \mathcal{U} - \{\hat{n}\}$, $N_{\text{assigned}} = |\hat{n}|$
   12. $\% N_{\text{assigned}} \leq \min\{|\mathcal{M}|, |\mathcal{U}|\}$
   13. for $i = 1$ to $N_{\text{assigned}}$
      14. $\hat{n} = \hat{n}(i)$, $\hat{m} = \hat{m}(i)$, $\hat{k} = \kappa_{\hat{n},\hat{m}}$
      15. $q^m_{\hat{k}} = (q^m_{\hat{k}} - [R_{m,k,n}T_i])^+$
   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 neither be forwarded to the UT nor will it be 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 following iterations and allocation instants. In [13], under uniform deployment of RSs and with different $M_{\text{const}}$, 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. So, besides 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. The computational complexity

The computational complexity of Variant-A and Variant-B schemes discussed in this paper is found to be polynomial in time of $O\left(\frac{N^2(N+M)}{M+1}\right)$, $M < N$, and $O\left(\frac{N^2(N+M+1)}{M+1}\right)$, $M+1 \leq N$, respectively. These complexity estimates come from the fact that the Hungarian algorithm is of $O\left(\binom{|M|}{2}\right)$. These are the complexity levels incurred in the second sub-frame. However, the proposed scheme inverts a slight increase in complexity of $O\left(\frac{N^2}{2}(K + M)\right)$ due to the first sub-frame allocation as compared to 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 3GPP LTE release 9 (Case 3) [26] or the WiMax Forum [27] while the WINNER C2 channel model [28] is used. In these system-level Matlab simulations, we have considered 19 hexagonal cells with 3 or 6 relays, with equal angular spacing, in each cell. The total DL frame length is 2 msec with equal sub-frame 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 for a Class 2 UT is $\lambda_2 = 2\lambda$ where $\lambda$ is 632 packets (188 bytes each) per second.

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 NLOS 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 omni-directional 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/hr, however the scheme can support mobility as high as 90 Km/hr, 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.

A. The HDR scheme vs. the quasi-FDR scheme and prior art

We first consider the case where all UTs belong to the same class, say Class 1, and the cellular network thus handles all symmetric traffic flows ($K_1 = K$). Figure 3 shows 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 are provided for all schemes, i.e., 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 (PFS), 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 inline with our understanding that the bottleneck results from the BS taking only a share of the resources to transmit directly 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 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 sub-frame.

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 5th

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS-RS distance</td>
<td>0.65 × cell radius</td>
</tr>
<tr>
<td>Minimum BS-UT distance</td>
<td>35 m</td>
</tr>
<tr>
<td>BS Tx. antenna gain</td>
<td>15 dB</td>
</tr>
<tr>
<td>RS Tx. antenna gain</td>
<td>10 dB</td>
</tr>
<tr>
<td>UT Rx. antenna gain</td>
<td>0 dB</td>
</tr>
<tr>
<td>No. of channel taps for BS-RS links</td>
<td>8</td>
</tr>
<tr>
<td>No. of channels taps for other links</td>
<td>6</td>
</tr>
<tr>
<td>TDD frame length</td>
<td>2 msec</td>
</tr>
<tr>
<td>Downlink : Uplink ratio</td>
<td>2:1</td>
</tr>
<tr>
<td>OFDM subcarrier bandwidth</td>
<td>10.9375 KHz</td>
</tr>
<tr>
<td>OFDM symbol duration</td>
<td>102.86 μsec</td>
</tr>
<tr>
<td>Subchannel width</td>
<td>18 subcarriers</td>
</tr>
<tr>
<td>CR-QAM target BER</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>Noise power density at Rx. nodes</td>
<td>-174 dBm/Hz</td>
</tr>
<tr>
<td>BS total Tx. power</td>
<td>46 dBm</td>
</tr>
<tr>
<td>RS total Tx. power</td>
<td>37 dBm</td>
</tr>
</tbody>
</table>

TABLE I SIMULATION PARAMETERS

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, Permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.
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 3 and 6 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) and thus improving the upper tail. However, the spatial diversity required to enhance the cell-edge throughput is lost and thus affecting the throughput in the long-term sense. This further underscores the superiority of the HDR scheme in highly loaded networks.

Figure 5 shows the average total cell throughput as a function of the number of UTs/cell. It is clear that the performance gap between the two variants of the HDR scheme and the quasi-FDR scheme increases significantly 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 one 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 sub-frame. 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 6 RSs is almost the same as that of Variant-A with 3 RSs, yet it is superior to that of the quasi-FDR with 6 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 a better performance exploiting the deployment geography; this is demonstrated here by the
top curve in this figure representing Variant-A with 6 RSs but using \( M_{\text{rsst}} = 2 \) closest RSs. As such, throughout the rest of our results, the proposed HDR scheme will be represented by Variant-A with \( M_{\text{rsst}} = 2 \). Figure 5 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 literature. A comparison with other non-relaying schemes can be found in [14] where we also show the latency improvement for relayed packets as compared to the quasi-FDR scheme based on Fig. 2.

B. The HDR Variant-A with symmetric and asymmetric traffic

We now consider the case where the UTs are equally divided into two groups \( K_1 \) and \( K_2 \), namely 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.

Figure 8 shows a scatter plot of user time-averaged throughput as a function of user location and shadowing with 20 UTs/cell with asymmetric traffic and 3 or 6 RSs. Other scatters are not shown for figure clarity.

relatively more spreading of the scatter points and less cell-edge throughput are realized for Class 2 UTs as compared to Class 1 UTs.

CDF plots of the scatter points with 6 and 3 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 especially 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 \( \beta \).

Comparing the CDF of Class 1 UTs in the case of asymmetric load \( (K = 20) \) to that of Class 1 in the reference case of symmetric load \( (K' = 30) \), it is observed that with 6 RSs, the
In general, all class-based absolute fairness curves are quite close while 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 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 6 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 details in [13]: 1- Implementing the constrained routing mode reduces the feedback overhead by a factor of $(M_{\text{rsd}} + 1)/M + 1$ 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 signalling overhead as compared to 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 $N_{\text{CSI}}/N^6$.

Since queue-awareness at the BS is a key element in the proposed RRA algorithms, it is important to investigate whether or not there is an associated overhead cost as compared to channel-aware only relay-based schemes. Looking at the queue length dynamics described in (12) to (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 investigation. As such, we observe that the BS can update the queue length information about its cell nodes spontaneously (at no cost), or at a minimal cost if system design necessitates, due to the following reasons:

1) Since UTs are the flow sinks of the DL traffic, their queue lengths are set to zero without incurring an overhead cost. Whereas the BS is self-aware of its full queue dynamics including the ARQ requests from the former recipients of its transmissions.

2) In contrast to 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 is used by the BS to increment the queue images of the destined RS(s).

\begin{align}
    x_j &= \frac{r_j}{\sum_{i=1}^{K} r_i / \beta_i} \\
    x_{w} &= \left(\frac{\sum_{i=1}^{K} r_{i,w} / \beta_i}{K \sum_{i=1}^{K} (r_{i,w} / \beta_i)^2}\right)^2
\end{align}

\(6\)Other results considering only the best 50% of link subchannels show no performance degradation.

---

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

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

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 to that in the all symmetric case, given the same aggregate load and the same resources. However, with 3 RSs and thus less spatial diversity, the improvement with $K' = 30$ becomes more visible.

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

\[ x_j = \frac{r_j}{\sum_{i=1}^{K} r_i / \beta_i} \]  

Similarly, the long-term fairness is shown in Fig. 11 using Jain’s index in (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.

\[ x_{w} = \left(\frac{\sum_{i=1}^{K} r_{i,w} / \beta_i}{K \sum_{i=1}^{K} (r_{i,w} / \beta_i)^2}\right)^2 \]
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.

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 sub-frame. 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. CONCLUSIONS

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 vs. the PFS. We present a novel throughput-optimal formulation in accordance with the emerging OFDMA-based cellular relay networks employing half-duplex relaying. Low-complexity iterative algorithms are devised to solve the formulated optimization over two consecutive sub-frames using the queue length coupling. Our numerical results show that with a slight complexity increase as compared to 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, besides the substantial improvement in both queue-awareness and latency. The results also 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 the time-average and long-term time scales. Load balancing among only active relays is jointly realized with the resource allocation.

IX. ACKNOWLEDGEMENT

The authors would like to thank Dr. Young-Doo Kim, Dr. Eungsun Kim, and Dr. Yoon-Chae Cheong, Samsung Electronics, SAIT, Korea, for the insightful discussions and invaluable support.

REFERENCES


Mohamed Rashad Salem (S’06) received his B.Sc. in Communications and Electronics from the department of Electrical Engineering, Alexandria University, Egypt, in 2000. He was nominated and hired as a faculty member in the department of Engineering Mathematics, Alexandria University, from which he received his M.Sc. degree and was promoted to the position of Assistant Lecturer in Feb 2006. Within this period he has gained wide experience in Research & Development and collaboration with industrial parties. In Jan 2011, Mr. Salem has earned the Ph.D. degree from the department of Systems and Computer Engineering, Carleton University, Ottawa, Canada. He has been conducting research in collaboration with Samsung Electronics on advanced radio resource management in next generation wireless networks. His research interests encompass stochastic modeling, congestion control, and optimization techniques. He has been granted the 2009/2010 Ontario Graduate Scholarship in Science and Technology (OGSST).

Abdulkareem Adinoyi obtained his Ph.D. degree from Carleton University, Canada, in 2006. He received the Masters degree from the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1998 and the B. Eng. degree from the University of Ilorin, Nigeria, in 1992. He has worked in the industry (as Engineer, Researcher, and Consultant) and University (as Lecturer and Assistant Professor). Dr. Adinoyi is an inventor of three patents (awarded and pending) in radio resource management for relay-based OFDMA networks. Between January 2004 and December 2006, he worked in the European Union 6th Framework integrated project - the WINNER. He is currently at Sweden Arabia as a consultant for Saudi Telecommunications Company. His research interests are in technology evolution, infrastructure-based multi-hop and relay networks, cooperative communication techniques and protocols, and radio resource management techniques for broadband wireless networks.

David Falconer received the B.A. Sc. degree in Engineering Physics from the University of Toronto in 1962, the S.M. and Ph.D. degrees in Electrical Engineering from M.I.T. in 1963 and 1967 respectively, and an honorary Doctorate of Science from the University of Edinburgh in 2009. After a year as a postdoctoral fellow at 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 1976-77 he was a visiting professor at Linköping University, Linköping, Sweden. Since 1980 he has been with Carleton University, Ottawa, Canada, where he is now Professor Emeritus and Distinguished Research Professor in the Department of Systems and Computer Engineering. His current research interests center around beyond-third-generation broadband wireless communications systems. He was Director of Carleton’s Broadband Communications and Wireless Systems (BCWS) 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 (WWRF) in 2004 and 2005. He received the 2008 Canadian award for Telecommunications Research, a 2008 IEEE Technical Committee for Wireless Communications Recognition Award, and the IEEE Canada 2009 Fessenden Award (Telecommunications). He is an IEEE Life Fellow.

Halim Yanikomeroglu received a B.Sc. degree in Electrical and Electronics Engineering from the Middle East Technical University, Ankara, Turkey, in 1990, and a M.A.Sc. degree in Electrical Engineering (now ECE) and a Ph.D. degree in Electrical and Computer Engineering from the University of Toronto, Canada, in 1992 and 1998, respectively. He was with the R&D Group of Marconi Komunikasyon A.S., Ankara, Turkey, from 1993 to 1994. Since 1998 Dr. Yanikomeroglu has been with the Department of Systems and Computer Engineering at Carleton University, Ottawa, where he is now an Associate Professor with tenure. Dr. Yanikomeroglu research interests cover many aspects of the physical, medium access, and networking layers of wireless communications with a special emphasis on multihop/relay/mesh networks and cooperative communications. Dr. Yanikomeroglu research is currently funded by Research In Motion (RIM, Canada), Huawei (China), Communications Research Centre of Canada (CRC), and NSERC. Dr. Yanikomeroglu is a recipient of the Carleton University Research Achievement Award 2009. Dr. Yanikomeroglu has been involved in the steering committees and technical program committees of numerous international conferences; he has also given 19 tutorials in such conferences. Dr. Yanikomeroglu is a member of the Steering Committee of the IEEE Wireless 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. Dr. Yanikomeroglu is the General Co-Chair of the IEEE Vehicular Technology Conference to be held in Ottawa in September 2010 (VTCF10-Fall). Dr. Yanikomeroglu was an editor for IEEE Transactions on Wireless Communications [2002-2005] and IEEE Communications Surveys & Tutorials [2002-2003], and a guest editor for Wiley Journal on Wireless Communications & Mobile Computing. He was an Officer of IEEE’s Technical Committee on Personal Communications (Chair: 2005-06, Vice-Chair: 2005-03, Secretary: 2001-02), and he was also a member of the IEEE Communications Society’s Technical Activities Council (2005-06). Dr. Yanikomeroglu is an Adjunct Professor at Prince Sultan Advanced Technologies Research Institute (PSATRI) at King Saud University, Riyadh, Saudi Arabia; he is also a registered Professional Engineer in the province of Ontario, Canada.