Exact Resource Allocation for Fair Wireless Relay

Edgar Arribas, Vincenzo Mancuso, Vicent Cholvi

***Abstract*—In relay-enabled cellular networks, the intertwined nature of network agents calls for complex schemes to allocate wireless resources. Resources need to be distributed among mobile users while considering how relay resources are allocated, and constrained by the traffic rate achievable by base stations and over backhaul links. In this letter, we derive an *exact* resource allocation scheme that achieves max–min fairness across mobile users, found with a linear complexity with respect to the number of mobile users *and* relays. The results reveal that the proposed scheme remarkably outperforms current solutions.**

***Index Terms*—Relay, fairness optimization, resource allocation.**

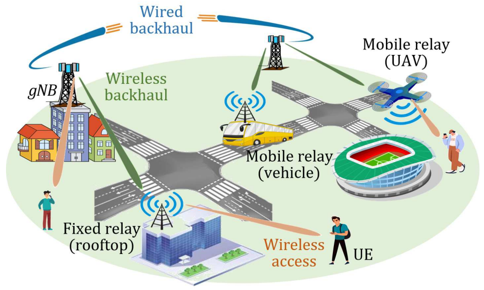

Figure 1: Reference scenario.

I. INTRODUCTION

We consider a heterogeneous relay-enabled network [1] formed by a set of fixed gNBs (*Next Generation Node B*) providing wireless service both to mobile users and relays. Figure 1 illustrates the considered scenario. It can be seen that there are two gNBs that provide service to one mobile user and three relays (a rooftop tower, a UAV (*Unmanned Aerial Vehicle*) and a bus). In turn, relays provide service to other mobile users (e.g., on the bus or in the stadium).

We derive a mechanism that provides a fair rate allocation to mobile users in downlink. Due to the high cost and limited availability of transmission resources, which are insufficient to accommodate all customers at peak quality, a meticulously planned allocation of resources is crucial to ensure user satisfaction in the provision of online services. Specifically, we guarantee $\max$–$\min$ fairness [1], i.e., we maximize the performance of the worst-case user, so potential service outages are minimized. Although alternative metrics exist for fairness, in this work we adopt $\max$–$\min$ because several practical systems require a minimum level of performance guarantees, below which the service cannot be properly deployed, hence customers would not pay for it. A wide range of services fall into this category: online streaming and real-time applications, augmented reality, etc. The quality of these services does not improve linearly or with a continuous function of, e.g., bandwidth and delay, but rather experiences a staircase quality function with very few steps, which saturates at some level [2]. For such services, what matters the most is to guarantee that all customers reach a level at which the service can be used.

The complexity of relay architectures makes the analysis quite difficult due to the intertwined nature of all the involved agents. Indeed, gNB resources must be allocated not only to directly served users, but also shared with relays, and relays may reuse wireless resources to serve their mobile users, thus generating interference. Additionally, the use of gNB resources is also constrained by the backhaul capacity. Finally, wireless resources must be assigned quickly to be able to adapt to changing scenarios, as guaranteed by our proposal.

Related Work

In the last years there has been an increasing number of studies focused on resource allocation in heterogeneous networks [1]. Although $\max$–$\min$ resource allocation for *single* cells was optimally resolved in [3], the extension of that problem to relay-aided networks is not trivial, and has been studied in different ways. Thus, here we review the available previous works, showing the different directions followed by them, and highlight how our work differs from existing results.

In [4], the authors focus on a downlink wireless network aided by a single UAV, which aims to maximize the minimum average rate among all users. In [5], the authors investigate the use of the non-orthogonal multiple access technique for the case of a single UAV relay and solve a joint channel-and-power allocation problem with an iterative algorithm under $\max$–$\min$ fairness, yet they do not achieve optimal results. Unlike our work, [4] and [5] do not consider the case of multiple relays.

In [6], the authors study proportional and $\max$–$\min$ fairness mechanisms in cognitive radio networks, where secondary users act as relays, aiming to provide acceptable rates. However, different from us, their analysis is restricted to Internet of Things scenarios and needs to solve non-convex problems, which prevents finding optimal results in reasonable time scales, while our approach finds exact solutions in linear time. In [7], the authors consider a scenario similar to ours. However, they take restrictive assumptions regarding how resources are allocated, and ignore inter-cell interference as well as interference between gNBs and relays. With that, they propose a suboptimal heuristic and show that it can improve fairness.

In [8], the authors consider satellite-terrestrial relay networks in which rates are maximized under fairness constraints for user association and spectrum allocation. However, the complexity leads them to resort to heuristics that are suboptimal, unless infinite iterations are run, which results

E. Arribas is with Universidad Cardenal Herrera-CEU, CEU Universities, València, Spain, edgar.arribasgimeno@uchceu.es. V. Mancuso is with IMDEA Networks Institute, Madrid, Spain,vincenzo.mancuso@imdea.org. V. Cholvi is with Universitat Jaume I (UJI), Castelló, Spain, vcholvi@uji.es.

Work supported by AEON-CPS (TSI-063000-2021-38), funded by the Ministry of Economic Affairs and Digital Transformation of Spain and the European Union NextGeneration-EU in the framework of the Spanish Recovery, Transformation and Resilience Plan; and by the grant INDI23/17 and GIR23/01 from Universidad Cardenal Herrera-CEU, CEU Universities.

impractical. In [9], authors address relay selection in dense heterogeneous networks to manage load balancing fairness, yet their focus is mainly oriented to device-to-device communications. However, with the approaches of [8] and [9], a minimum service level for users cannot be guaranteed, different from what addressed in our work.

Available works differ from our proposal in the sense that they either use just one relay, address different communication scenarios, or approach fairness in ways that cannot guarantee a minimum service performance, all of them ignoring in fact the presence of backhaul bottlenecks.

Contributions

Novelty and contributions of this letter are as follows:

- We develop a max–min fair resource allocation scheme for wireless relay networks that allows to jointly allocate resources to both mobile users *and* several relays, considering wired and wireless backhaul bottleneck constraints, which precludes the direct use of existing schedulers.
- Our algorithm finds the *exact* solution for the associated optimization, which goes beyond existing results.
- Such exact solution is found with linear complexity on the number of mobile users *and* relays, which is a strong advantage when it comes to practical implementations.
- The performance evaluation shows that our proposal remarkably outperforms current schemes when adapted to the framework of wireless relay networks, revealing that, actually, the scheme derived in this letter is needed.

II. SYSTEM MODEL

Table I summarizes the system model parameters used in this letter. We consider a wireless relay-enabled network composed by a set of fixed *gNB*s and a set of relays that provide cellular service to a set of mobile users. We model downlink traffic, i.e, traffic eventually delivered to mobile users by a *gNB* or a relay. Each *gNB* is attached to a wired backhaul network, whereas each relay is attached to one *gNB* by means of a wireless backhaul link. This represents a realistic framework for heterogeneous cellular networks that offers (i) a flexible way to adapt to occasional events and emergencies (e.g., from the case of crowded events to the case in which cellular coverage has to be temporary brought where no coverage is typically needed or because of an emergency or a specific "mission" requiring network support upgrades) and (ii) an affordable way to extend network services without incurring the costs of a fixed infrastructure extension (e.g., when a "volatile" infrastructure is needed and the cost of a fixed one would not be otherwise recovered through the revenue associated with the service) [1].

The set of relays attached to *gNB* g is denoted as $\mathcal{R}_g$, the set of *gNB*-served users is denoted as $\mathcal{U}_g$, for each *gNB* g, the set of users served by relay r is denoted as $\mathcal{U}_r$, and the set of users served by some relay attached to *gNB* g is denoted as $\mathcal{U}_g^*$.

Each *gNB* g receives a maximum traffic capacity rate (denoted τ_g) from the wired backhaul network, perhaps different from that of the other *gNB*s. We denote as W_{relays}^g the bandwidth of *gNB* g dedicated to relays and as W_{users}^g the bandwidth of *gNB* g dedicated to users directly attached to g. In addition, each relay r will allocate its bandwidth, which we denote as W_{users}^r, among the users it serves (note that W_{relays}^g, W_{users}^g and W_{users}^r are fixed values, since the assignment of spectrum bands to operators is performed by means of government auctions where only channels of fixed bandwidth are offered [10]). Such bands for mobile users and relays may be deployed by the operator as either orthogonal or reused bands. What matters for our analysis is that interference, if present, is accounted for. After that, operators can split the assigned bandwidth into smaller portions to allocate subchannels to specific groups of users and services, according to their target (e.g., optimize a fair network performance).

TABLE I: SYSTEM MODEL PARAMETERS

Parameter	*Description*
$\mathcal{R}_g$	Set of relays attached to *gNB* g.
$\mathcal{U}_g, \mathcal{U}_r, \mathcal{U}_g^*$	Mobile users attached to *gNB* g, mobile users attached to a relay r, and mobile users attached to any relay r (i.e., $\mathcal{U}_g^* = \bigcup_{r \in \mathcal{R}_g} \mathcal{U}_r$).
τ_g	Maximum traffic rate of *gNB* g.
W_{relays}^g, W_{users}^g, W_{users}^r	Bandwidth of *gNB* g dedicated to relays, bandwidth of *gNB* g dedicated to mobile users and bandwidth of relay r (dedicated to mobile users).
$W_{\text{relays}}^{\min}$, $W_{\text{users}}^{\min}$	Minimum bandwidth for each relay and mobile user.
$\gamma_{s,y}$	SINR between s and y, where s is a station (a *gNB* or a relay) and y is either a mobile user or a relay.

On another hand, it must be taken into account that practical systems cannot assign arbitrarily small bandwidth to individual stations or users [11]. Concretely, each relay obtains *at least* $W_{\text{relays}}^{\min}$, while each served mobile user receives *at least* $W_{\text{users}}^{\min}$.

Mobile users access downlink wireless resources with an OFDMA (*Orthogonal Frequency-Division Multiple Access*) scheme, as for 3GPP mobile broadband networks [10], which enables multiple devices to exchange data concurrently over a shared frequency band by dividing it into orthogonal subcarriers. We assume that all *gNB*s and relays use their entire available bandwidth, which in practice, is the case that requires optimization. Hence, we consider that mean SINR (*Signal to Interference and Noise Ratio*) values are constant with respect to user resource allocation, and are solely determined by the inter-cell interference level, which in turn depends on which frequencies are used by *gNB*s and relays. Instead, scheduling at the *gNB* or relay prevents intra-cell interference.

Although the above-mentioned interference can be reduced by making *gNB*s use 3D-beamforming or adopting orthogonal frequencies, depending on the scenario it will be necessary to take into account the signal strength of each wireless channel, measured as the SINR. We denote by $\gamma_{g,r}$ the SINR of the relay link between *gNB* g and a relay r, and by $\gamma_{s,u}$ the SINR of the access link between a station s (either a *gNB* or a relay) and a mobile user u. As wireless networks perform resource allocation based on the channel state information perceived (basically, the SINR observed), at the moment of distributing resources the scheduler is already aware of the users and relays cell selection and thus the SINR channel values, so that those γ parameters need to be considered here as problem inputs.

III. THE RESOURCE ALLOCATION PROBLEM

The aim of our work is to optimize the max-min fairness of the throughput received by mobile users. This is not a trivial task, as all the involved agents (*gNB*s, relays and mobile users) are intertwined (e.g., resources of mobile users from one relay cannot be allocated without knowing what backhaul resources that relay will get, depending on other relay resources and the *gNB* bottleneck over the wired backhaul), while the interference management also involves different types of colliding wireless channels. Since at resource allocation the network disposes of the CSI (*Channel State Information*) feedback necessary to know the SINRs of the channels, each *gNB* will be able to solve the resource allocation problem for its relays, its mobile users, and the users of relays attached to that *gNB* in a concurrent and independent manner, by using the convex program that we will introduce next in (1).

More formally, it will be necessary to obtain, for each relay r and for each mobile user u, both the share of bandwidth assigned (denoted w_r and w_u), and the throughput experienced by the network node (denoted T_r and T_u).

In (1) we formulate, for each *gNB* g, the corresponding resource allocation optimization in a convex program:

$$\begin{cases} \quad \max\min\left\{T_u \,|\, u\in\mathcal{U}_g \bigcup \mathcal{U}_g^*\right\}, \quad \text{s.t.:} \\ \\ 1.\ w_r \geq W_{\text{relays}}^{\min}, & \forall r\in\mathcal{R}_g; \\ 2.\ \sum_{r\in\mathcal{R}_g} w_r = W_{\text{relays}}^{g}; \\ 3.\ T_r \leq w_r \log_2(1+\gamma_{g,r}), & \forall r\in\mathcal{R}_g; \\ \\ 4.\ w_u \geq W_{\text{users}}^{\min}, & \forall u\in\mathcal{U}_g \bigcup \mathcal{U}_g^*; \\ 5.\ \sum_{u\in\mathcal{U}_g} w_u = W_{\text{users}}^{g}; \\ 6.\ \sum_{u\in\mathcal{U}_r} w_u = W_{\text{users}}^{r}, & \forall r\in\mathcal{R}_g; \\ \\ 7.\ T_u \leq w_u \log_2(1+\gamma_{s,u}), & \forall (s,u)\in(\{g\}\bigcup\mathcal{R}_g)\times(\mathcal{U}_g\bigcup\mathcal{U}_g^*); \\ 8.\ \sum_{u\in\mathcal{U}_r} T_u \leq T_r, & \forall r\in\mathcal{R}_g; \\ 9.\ \sum\limits_{u\in\mathcal{U}_g} T_u + \sum\limits_{r\in\mathcal{R}_g} T_r \leq \tau_g. \end{cases} \tag{1}$$

The first three constraints are related to the backhaul. The first guarantees that each relay obtains a minimum bandwidth, the second states that the aggregated bandwidth of relays is fixed, and the third is Shannon capacity.

The fourth constraint guarantees a minimum bandwidth for each served user, while the fifth and sixth constraints state that the aggregate share bandwidth of these users must adjust to the whole channel capacity allowed by their serving station.

The seventh constraint restricts the throughput allocated to mobile users to the Shannon capacity. The eighth constraint expresses the fact that the throughput allocated to relay-served users cannot exceed the wireless backhaul capacity assigned to the relay. Finally, the ninth constraint states that the aggregate throughput served by a *gNB* (to mobile users *and* relays) cannot exceed the *gNB* bottleneck over the wired backhaul.

The optimization program in (1) is convex, hence solvable in polynomial time with standard interior-point methods [12]. Yet, such methods have a cubic computational complexity with respect to the number of mobile users [13], which is prohibitive for real-time applications with large mobile user populations. Thus, in the next section, we derive an exact analytical solution that has a linear complexity with respect to the number of mobile users *and* relays attached to the *gNB*.

IV. THE EXACT max–min RESOURCE ALLOCATION

In this section, we introduce `LinEx`: a scheme that provides, in *linear* time, the *exact* max–min resource allocation for the type of wireless relay networks described in Section II.

Algorithm 1 `LinEx`: The *linear* and *exact* max–min allocation.

```
1:  Start: gNB $g$, $w_r \leftarrow W_{\text{relays}}^{\min}$ and $T_r \leftarrow w_r\log_2(1+\gamma_{g,r})$, $\forall r\in\mathcal{R}_g$.
2:  Derive the optimal rates $\{T_u\}_{u\in\mathcal{U}_g\cup\,\mathcal{U}_g^*}$ for all users, limited to
    the wireless relay traffic of $T_r$ and ignoring the wired bottleneck.
3:  $\beta \leftarrow 1$.
4:  while $\beta = 1$ do
5:      $T_m \leftarrow \min\{T_u \,|\, T_u < w_u\log_2(1+\gamma_{r,u}),\, r\in\mathcal{R}_g, u\in\mathcal{U}_r\}$.
6:      $\mathcal{L}_m \leftarrow \{u\in\mathcal{U}_g^* \,\mid\, T_u = T_m\}$.
7:      $T_M \leftarrow \min\{T_u \,\mid\, T_u > T_m, u\in\mathcal{U}_g^*\}$.
8:      $T_{M_2} \leftarrow \min(T_M, \min\{w_u\log_2(1+\gamma_{r,u}) \,|\, r\in\mathcal{R}_g, u\in\mathcal{L}_m\cap\mathcal{U}_r\})$.
9:      $\overline{\mathcal{U}}_r \leftarrow \{u\in\mathcal{U}_r \,\mid\, u\in\mathcal{L}_m\}$, $\forall r\in\mathcal{R}_g$.
10:     $\beta \leftarrow \min\left(1, \frac{W_{\text{relays}}^{g} - \sum_{r\in\mathcal{R}_g} w_r}{(T_{M_2}-T_m)\cdot\sum_{r\in\mathcal{R}_g}|\overline{\mathcal{U}}_r|/\log_2(1+\gamma_{g,r})}\right)$.
11:     $w_r \leftarrow w_r + |\overline{\mathcal{U}}_r|\beta\,(T_{M_2}-T_m)\,/\log_2(1+\gamma_{g,r})$, $\forall r\in\mathcal{R}_g$.
12:     $T_r \leftarrow w_r\log_2(1+\gamma_{g,r})$, $\forall r\in\mathcal{R}_g$.
13:     $T_u \leftarrow T_u + \beta\,(T_{M_2}-T_m)$, $\forall u\in\mathcal{L}_m$.
14: end while
15: $T_r \leftarrow \sum_{u\in\mathcal{U}_r} T_u$, $\forall r\in\mathcal{R}_g$.
16: if $\sum_{u\in\mathcal{U}_g} T_u + \sum_{r\in\mathcal{R}_g} T_r > \tau_g$ then
17:     reduce the rates starting from the highest until the constraint
        on $\tau_g$ in (1) is satisfied, preserving max–min fairness.
18: end if
```

The `LinEx` scheme (cf. Algorithm 1) is independently executed at each *gNB* g, and works as follows:

- First of all, it assigns the minimum bandwidth $w_r = W_{\text{relays}}^{\min}$ and the highest achievable rate $T_r = w_r\log_2(1+\gamma_{g,r})$ to each relay $r\in\mathcal{R}_g$ (cf. Step 1).
- Then, it derives the optimal rates for all the users directly attached to either g or to relays, limited to the relay backhaul traffic of T_r and ignoring the wired bottleneck. Such subproblem has similarities with the one studied in [3], whose solution is well known (for the interested reader, more details are provided in a separate technical report [14], where we show how this solution can be adapted to our system with one bottleneck).
- Now, we increase as much as possible the utilities by equally raising the lowest values of $\{T_u\}_{u\in\mathcal{U}_r}$, $\forall r\in\mathcal{R}_g$ (as long as constraints are not violated). Let

  $$T_m = \min_{u\in\mathcal{U}_r}\{T_u \,|\, T_u < w_u\log_2(1+\gamma_{r,u}),\, r\in\mathcal{R}_g\} \tag{2}$$

  be the minimum throughput rate that has not reached Shannon capacity (if T_m does not exist, we are done). Let

  $$\mathcal{L}_m = \{u\in\mathcal{U}_g^* \,\mid\, T_u = T_m\} \tag{3}$$

  be the set of those relay–served users such that their rate is the same as the minimum T_m. Let

  $$T_M = \min\{T_u \,\mid\, T_u > T_m, u\in\mathcal{U}_g^*\} \tag{4}$$

  be the minimum rate among relay–served user rates that are not as the minimum T_m (cf. step 7). Let's further refine such minimum by considering the Shannon capacity of users in $\mathcal{L}_m$, which are in the worst serving condition:

  $$T_{M_2} = \min\left(T_M, \min_{r\in\mathcal{R}_g}\{w_u\log_2(1+\gamma_{r,u}) \,|\, u\in\mathcal{L}_m\cap\mathcal{U}_r\}\right). \tag{5}$$

  The goal now is to increase $\{T_u\}_{u\in\mathcal{L}_m}$ as much as possible, without exceeding T_{M_2}, as long as those involved relays $r\in\mathcal{R}_g$ can request more resources to increase T_r. Let

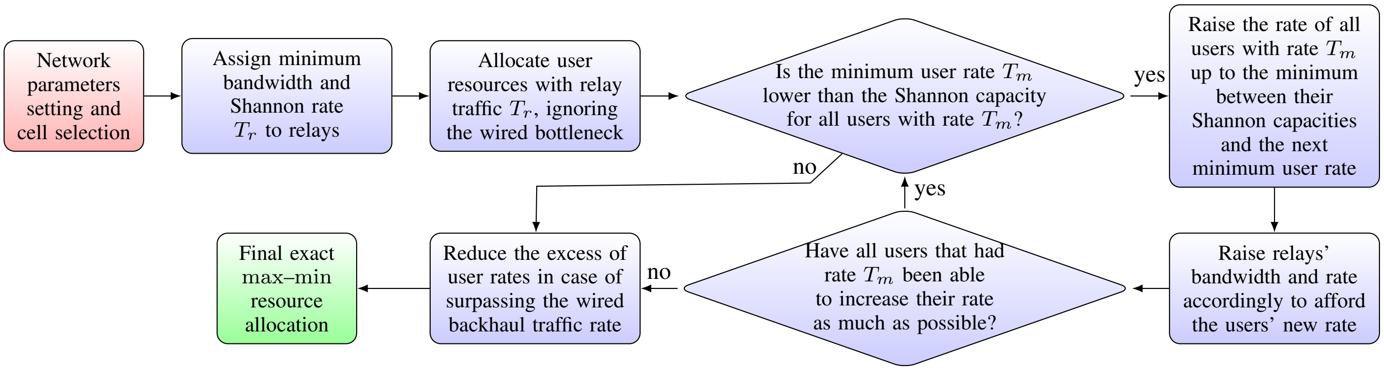

Figure 2: Flowchart diagram of the `LinEx` scheme operation.

$\beta \in [0,1]$ be an auxiliary parameter that we will better define later. $\{T_u\}_{u\in\mathcal{L}_m}$ will be increased by $\beta(T_{M_2}-T_m)$, i.e., at most, by $T_{M_2}-T_m$ (cf. step 13). Let

$$\overline{\mathcal{U}}_r = \{u \in \mathcal{U}_r \mid u \in \mathcal{L}_m\}, \quad \forall r \in \mathcal{R}_g. \tag{6}$$

Now, we set $T_u^{'} = T_u + \beta(T_{M_2}-T_m)$, $\forall u \in \mathcal{L}_m$ to increase the involved throughput rates. Hence, we set $\forall r \in \mathcal{R}_g$:

$$\begin{aligned} T_r &= \sum\nolimits_{u\notin\overline{\mathcal{U}}_r} T_u + \sum\nolimits_{u\in\overline{\mathcal{U}}_r} T_u^{'} \\ &= \sum\nolimits_{u\notin\overline{\mathcal{U}}_r} T_u + \sum\nolimits_{u\in\overline{\mathcal{U}}_r} (T_u + \beta\,(T_{M_2}-T_m)) \\ &= \sum\nolimits_{u\notin\overline{\mathcal{U}}_r} T_u + \sum\nolimits_{u\in\overline{\mathcal{U}}_r} T_u + |\overline{\mathcal{U}}_r|\beta(T_{M_2}-T_m). \end{aligned} \tag{7}$$

Hence, in step 11 we set $\forall r \in \mathcal{R}_g$:

$$\begin{aligned} w_r^{new} &= \frac{T_r}{\log_2(1+\gamma_{g,r})} = \frac{\sum_{u\in\mathcal{U}_r} T_u + |\overline{\mathcal{U}}_r|\beta(T_{M_2}-T_m)}{\log_2(1+\gamma_{g,r})} \\ &= w_r + \frac{|\overline{\mathcal{U}}_r|\beta(T_{M_2}-T_m)}{\log_2(1+\gamma_{g,r})}. \end{aligned} \tag{8}$$

The aggregation of the new relay resource allocation has to be lower than the total bandwidth, i.e.,

$$\begin{aligned} \sum\nolimits_{r\in\mathcal{R}_g} w_r^{new} &= \sum\nolimits_{r\in\mathcal{R}_g} \left(w_r + \frac{|\overline{\mathcal{U}}_r|\beta\,(T_{M_2}-T_m)}{\log_2(1+\gamma_{g,r})} \right) \\ = \sum\nolimits_{r\in\mathcal{R}_g} w_r &+ \beta(T_{M_2}-T_m) \sum\nolimits_{r\in\mathcal{R}_g} \frac{|\overline{\mathcal{U}}_r|}{\log_2(1+\gamma_{g,r})} \end{aligned} \tag{9}$$

has to be lower than or equal to W_{relays}^{g}. Hence, isolating β we get that necessarily:

$$\beta \leq \frac{W_{\text{relays}}^{g} - \sum_{r\in\mathcal{R}_g} w_r}{(T_{M_2}-T_m)\sum_{r\in\mathcal{R}_g} |\overline{\mathcal{U}}_r| / \log_2(1+\gamma_{g,r})}. \tag{10}$$

Hence, in step 10 we have defined β as:

$$\beta = \min\left(1, \frac{W_{\text{relays}}^{g} - \sum_{r\in\mathcal{R}_g} w_r}{(T_{M_2}-T_m)\sum_{r\in\mathcal{R}_g} |\overline{\mathcal{U}}_r| / \log_2(1+\gamma_{g,r})}\right). \tag{11}$$

Once the parameter β is derived, we assign $w_r = w_r^{new}$ and $T_r = w_r \log_2(1+\gamma_{g,r})$, $\forall r \in \mathcal{R}_g$ (cf. step 12). In the case that $\beta = 1$ (cf. step 4), we repeat the process defining T_m again and increasing the corresponding throughput rates.

- To finalize the allocation and guarantee an exact solution, we need to ensure that the constraint on τ_g in (1) holds, which is done in steps 16–18. Whereas such reduction can be performed in a number of ways, in [14] we provide an algorithm preserves max–min fairness: it reduces user throughputs from above $T = \min\{T_u\}$ down to T, at most, starting from the highest one, so that the aggregated network throughput reaches τ_g; and if that is not enough, then assigns $T_u = \tau_g / |\mathcal{U}_g \cup \mathcal{U}_g^*|$.

For a better understanding of the `LinEx` scheme, in Figure 2 we show a flowchart with a summary of the `LinEx` operation.

Computational Complexity Analysis

The `LinEx` scheme guarantees the exact max-min fairness. However, it is important to ensure that the proposed solution is deployable. Indeed, the `LinEx` scheme has a linear complexity in the number of the operations with respect to the number of mobile users *and* the number of relays (i.e., the complexity is $\mathcal{O}\left(|\mathcal{U}_g \bigcup \mathcal{U}_g^*| \cdot |\mathcal{R}_g|\right)$, for each $gNB\ g$), as shown next.

In Algorithm 1, the initial stage of deriving the user rates ignoring bottlenecks is solved in linear time with water-filling schemes [3]. Then, the while loop will run over, at most, as many iterations as the number of relay-served mobile users. That happens because the while loop stops when $\beta < 1$. However, that only happens when there are not enough resources to increase the resources for mobile users gathered in $\mathcal{L}_m$ (which grows, at least, by one mobile user at each iteration). Then, within the loop, we compute sums over the number of relays (i.e., $|\mathcal{R}_g|$), as we thoroughly detail in a technical report [14]. Afterwards, we sum the user rates for each relay and, finally, the excess of throughputs is optimally reduced to meet the wired bottleneck constraint with a linear descendent search. Hence, the overall complexity of the `LinEx` scheme in Algorithm 1 is $\mathcal{O}\left(|\mathcal{U}_g \bigcup \mathcal{U}_g^*| \cdot |\mathcal{R}_g|\right)$.

V. PERFORMANCE EVALUATION

Here we present a performance evaluation of the `LinEx` scheme. For that, we compare our proposal with two benchmarking schemes: the `CSolver` and the `WFill` schemes.

On the one hand, `CSolver` consists of a *convex* optimization *solver* that provides optimal solutions. Such optimizer has a high complexity (of the *cubic* order) that makes it undeployable in practice. However, it will allow us to verify that, indeed, our scheme provides optimal solutions.

On the other hand, the `WFill` scheme implements the solution of the max–min resource allocation problem based on the known legacy allocation in [3], following *water-filling* algorithms. Such a solution has been shown to be optimal when base stations are considered individually, yet it does

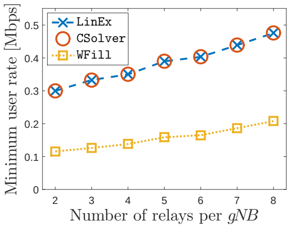

Figure 3: Wireless relay network with 3 *gNB*s and $U = 600$ users.

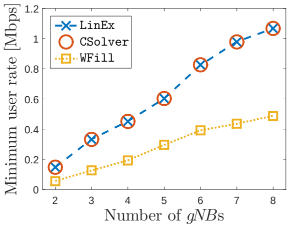

Figure 4: Wireless relay network with 3 relays per *gNB* and $U = 600$.

not take into account the interwined nature of multiple-source allocations jointly constrained by wireless and wired bottlenecks. That is the main difference between the `WFill` and the `LinEx` schemes: the former is the result of adapting the legacy scheduler to wireless relay networks, while the latter has been thoughtfully designed to take into account the backhaul resources and traffic constraints.

All simulations are run over uniformly random network topologies in a circular region with radius of 750 m. Relays are considered as aerial relays, so that all network parameters and channel models are taken as in the realistic environment of [15]: a heterogeneous dense urban network with terrestrial path-loss models and aerial channel fading based on LoS (*Line-of-Sight*) communications for relay-served users (for the interested reader, more details are provided in [14]). The carrier frequency for *gNB*s is 1815.1 MHz both for wireless backhaul (for transmissions to relays) and access channels (to mobile users), while for relays the carrier is 2630 MHz, with 20 MHz of band in all cases. Transmissions from *gNB*s to relays do not interfere with transmissions from *gNB*s to mobile users on the ground thanks to the adoption of precise 3D-beamforming over clear LoS links to the aerial relays. Results are averaged over 1000 runs.

In Figure 3 and Figure 4 we observe the utility achieved (i.e., the minimum user rate) in two cases: (i) when we increase the number of relays served by each *gNB* in a network with 3 *gNB*s and (ii) when we increase the number of *gNB*s, each *gNB* serving 3 relays. In both cases there are $U = 600$ mobile users that attach to the *gNB* or relay cell with strongest signal (as in the operational 3GPP networks) and the wired bottleneck traffic is of $\tau_g = 180$ Mbps for each *gNB* g.

Firstly, we see that the `LinEx` and `CSolver` schemes perform equally in all cases. That means that `LinEx` finds always the optimal $\max$–$\min$ resource allocation, with the important difference that `LinEx` finds it in linear time, while the complexity of `CSolver` is, instead, of the cubic order. Secondly, we observe that as long as relays or *gNB*s are added, network performance clearly increases. Indeed, the minimum user rate increases as users can find better connections and resource splitting opportunities. Finally and most importantly, we remark that the performance of `WFill` is between 30% to 60% worse than `LinEx`. This shows that not only `LinEx` is linear and exact, but it also considerably outperforms available state-of-the-art proposals. Such a result reveals that it becomes crucial to account for the intertwined nature of multiple resource allocation at different cells, altogether constrained by backhaul resources and traffic rates. Instead, simply adapting available allocation schemes to the wireless relay context is insufficient to achieve an acceptable network performance. In conclusion, `LinEx` stands as an efficient and lightweight implementable scheme for $\max$–$\min$ fair resource allocation in current wireless relay networks.

VI. CONCLUSIONS

We have solved the optimal $\max$–$\min$ allocation of downlink resources in wireless relay-enabled networks. With `LinEx`, the proposed exact $\max$–$\min$ resource allocation scheme, we have shown that the optimal distribution of resources can be found in linear time on the number of mobile users *and* relays, which is a key enabler for implementation over cellular networks. Considering backhaul bottlenecks result to be crucial to assign resources to mobile users depending on the allocation to other relays and users. We have shown that not only our algorithm finds the optimal performance in terms of $\max$–$\min$ fairness in linear time, but it also stands as the only practical solution to enable $\max$–$\min$ fair resource allocation in wireless relay networks.

REFERENCES

[1] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, “A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges,” *IEEE Communications Surveys and Tutorials*, vol. 23, no. 2, pp. 668–695, 2021.

[2] S. S. Sabet, S. Schmidt, S. Zadtootaghaj, B. Naderi, C. Griwodz, and S. Möller, “A latency compensation technique based on game characteristics to mitigate the influence of delay on cloud gaming quality of experience,” in *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys ’20. ACM, 2020, p. 15–25.

[3] A. Coluccia, A. D’Alconzo, and F. Ricciato, “On the optimality of max–min fairness in resource allocation,” *annals of telecommunications-annales des télécommunications*, vol. 67, pp. 15–26, 2012.

[4] Y. Guo, S. Yin, and J. Hao, “Resource allocation and 3-D trajectory design in wireless networks assisted by rechargeable UAV,” *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 781–784, 2019.

[5] D. Zhai, H. Li, X. Tang, R. Zhang, Z. Ding, and F. R. Yu, “Height optimization and resource allocation for NOMA enhanced UAV-aided relay networks,” *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 962–975, 2021.

[6] N. S. Moayedian, S. Salehi, and M. Khabbazian, “Fair resource allocation in cooperative cognitive radio IOT networks,” *IEEE Access*, vol. 8, pp. 191 067–191 079, 2020.

[7] O. Elgendy, M. Ismail, and K. Elsayed, “Radio resource management for LTE-A relay-enhanced cells with spatial reuse and max-min fairness,” *Telecommunication Systems*, vol. 68, no. 4, pp. 643–655, 2018.

[8] Z. Tariq, H. Z. Khan, U. Fakhar, M. Ali, A. N. Akhtar, M. Naeem, and A. Wakeel, “Fairness-based user association and resource blocks allocation in satellite–terrestrial integrated networks,” *Physical Communication*, vol. 55, p. 101934, 2022.

[9] N. Moghaddas-Gholian, V. Solouk, and H. Kalbkhani, “Relay selection and power allocation for energy-load efficient network-coded cooperative unicast D2D communications,” *Peer-to-Peer Networking and Applications*, vol. 15, no. 2, pp. 1281–1293, 2022.

[10] 3GPP, “Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception,” accessed: 2023-06-21.

[11] Y. Liu, *Introduction to OFDM receiver design and simulation*. Artech House, 2019.

[12] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. Siam, 1994, vol. 13.

[13] S. Bubeck *et al.*, “Convex optimization: Algorithms and complexity,” *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[14] E. Arribas, V. Cholvi, and V. Mancuso, “Exact resource allocation for fair wireless relay,” *arXiv preprint arXiv:2307.06095*, 2023.

[15] E. Arribas, V. Mancuso, and V. Cholvi, “Coverage optimization with a dynamic network of drone relays,” *IEEE Transactions on Mobile Computing*, vol. 19, no. 10, pp. 2278–2298, 2019.