EURECOM
Department of Communication Systems
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report 16-328

HoP: Hierarchize then Optimize: A distributed framework for User Association and Flexible TDD Allocation for Access and Backhaul Networks

December 2016

Nikolaos Sapountzis, Thrasyvoulos Spyropoulos, Navid Nikaein and Umer Salim.

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {Nikolaos.Sapountzis, Thrasyvoulos.Spyropoulos, Navid.Nikaein }@eurecom.fr, Umer.Salim@intel.com

[1]EURECOM's research is partially supported by its industrial members: BMW Group Research and Technology, IABG, Monaco Telecom, Orange, Principaut de Monaco, SAP, ST Microelectronics, Symantec.

HoP: Hierarchize then Optimize: A distributed framework for User Association and Flexible TDD Allocation for Access and Backhaul Networks

Nikolaos Sapountzis, Thrasyvoulos Spyropoulos, Navid Nikaein and Umer Salim.

Abstract

The success of future heterogeneous networks (HetNets) heavily depends on the interplay between user association and resource allocation on both the access and backhaul network. While user association is key to improve both the user and network performance, it is becoming a multi-objective optimization problem that should consider the number and type of BS in range. Furthermore, the increasing spatio-temporal heterogeneity in downlink(DL) and uplink(UL) traffic suggests that DL/UL resources can be tuned to optimally serve the respective workload. Split DL/UL association and flexible TDD offer such an opportunity. While much literature exists on these problems, the majority consider them separately. In this work, we develop a framework that tackles the optimal interplay of (i) user-association, (ii) radio resource allocation, and (iii) backhaul resource allocation of TDD resources, for a family of objective functions. We propose an algorithm that reduces the complexity of this problem by decomposing it into three optimization subproblems, each potentially solved by a different network element and at different timescales. We prove convergence to the global optimum, and provide simulation results that demonstrate the performance benefits of our approach.

Index Terms

user association, backhaul, queueing theory, uplink, downlink, hetnets, resource allocation, dynamic TDD.

Contents

List of Figures

1 Introduction

Lately, heterogeneous network (HetNet) deployments have been widely considered in 4G and beyond wireless networks. They are composed of conventional macro cells (MC) overlaid with a set of low-power small cells (SC). Due to the increasing number and type of base stations (BS) within the range of each user, the problem of user association becomes increasingly important. More advanced schemes beyond simple SINR-based ones are thus needed [1, 2] to balance user- and network-related performance goals.

While optimization of most current networks revolves around the downlink (DL) performance, social networks, augmented reality games, and other uplink (UL)-intensive envisioned applications suggest that UL performance becomes as important. Recent approaches that aim to improve both DL and UL throughput suggest that UL/DL association should be in fact decoupled for optimal performance. As one example, a user equipment (UE) could be connected to a macro BS in the DL (from which it receives the highest signal level), and to an SC in the UL (where the pathloss is lower) [3, 4]. However, if the DL resources of the macro BS, or the UL resources of the SC are not sufficient, this approach can lead to *unnecessary* congestion or under-utilization in either direction.

Typically, in today's systems, each BS is given an amount of bandwidth resources to utilize for both DL and UL traffic by duplexing on the frequency (Frequency Division Duplex-FDD) or the time (Time Division Duplex-TDD) domain. While conventional networks are mainly designed for FDD or pre-configured TDD schemes, heterogeneous traffic demand, desired architectural flexibility, and scarcity of spectrum has increased interest in *flexible TDD* schemes, that can *match the UL and DL resources to the actual demand* [5].

Nevertheless, dynamic/flexible TDD schemes require additional considerations, in particular in assymetric interference scenarios. As a typical example, if an SC is doing UL while a nearby MC is transmitting on the DL (with much higher power), the performance of the SC might be significantly degraded from this *cross-interference*. Enhanced Inter-Cell Interference Coordination (eICIC) schemes such as Almost Blank Subframes (ABS) could alleviate this but only to some extent [6, 7]. Large amounts of mismatch might lead to excessive usage of resources for eICIC, instead of user traffic, leading instead to considerable performance degradation. Many additional allocation schemes have further been proposed to tackle this problem(s) [8] [9] [10], most of them revolving around a key-enabler for 5G networks, namely "enhanced Interference Mitigation and Traffic Adaptation" (eIMTA), standardized in LTE-A Release 13 [11]. However, it is not clear which scheme is the best option and how it should interact with user association.

Finally, a common limitation of most of the above works is that they focus solely on the radio access part, ignoring the backhaul (BH) network. This might be reasonable for legacy cellular networks, given that the macro-cell backhaul is often over-provisioned (e.g., fiber). However, expected backhaul limitations for small cells [12] and the additional backhaul load for coordinated transmission (CoMP)

and eICIC put a heavy toll on backhaul links, that might become the new bottleneck. This calls for a joint optimization of radio and backhaul [13–15]. Nevertheless, these works mostly focus on the DL [13, 14, 16]. A recent work [15] analytically derives jointly optimal UL and DL user association rules for various backhaul-limited scenarios. However, their work assumes fixed amount of resources, pre-allocated for UL and DL, for both the radio access and backhaul. Interestingly, the authors there show that pre-configured backhaul resource allocation further penalizes performance. Undoubtedly, backhaul resource allocation policies should interact with the *user association* and *flexible TDD* radio access policies, in order to satisfy the UL and DL traffic demands that the latter generate.

In this paper, we propose an optimization framework that jointly considers all these problem dimensions. To our best knowledge, this is the first work to attempt it. Our main contributions can be summarized as follows:
(1) We propose an analytical framework to study the interplay between (i) user association, (i) radio access resource allocation with cross-interference management, and (iii) backhaul resource allocation, significantly extending the popular framework of [1]. (Section 2)
(2) We show that the joint problem is non-convex, unlike variants studied in the past [1, 3, 15, 16], but posseses some "hidden" convexity properties that allows its decomposition into three subproblems. These subproblems can be solved through convex optimizers, at possibly different elements (e.g. UE, BS, backhaul link), and at different timescales, facilitating a hierarchical implementation. (Section 3)
(3) Using extensive simulations, we highlight complex trade-offs involved between the different subproblems, and show that significant performance improvements could be achieved compared to current standards. (Section 4)

2 System Model and Assumptions

We use a similar problem setup as the one used in a number of related works [1, 3, 15, 17], and extend it accordingly. To keep notation consistent, for all variables considered, the superscript "D" and "U" refer to downlink and uplink traffic, respectively. For brevity, in the following *we present most notation and assumptions in terms of downlink traffic only, assuming that the uplink case and notation is symmetric*. Specific differences will be elaborated, where necessary. In Table 1, we summarize some useful notation.

2.1 Traffic Model

(A.1 - Traffic arrival rates) Traffic at location $x \in \mathcal{L}$ consists of file (or more generally *flow*) requests arriving according to an inhomogeneous Poisson point

Table 1: Notation

	Downlink	**Uplink**
Access Resource Allocation Policy for BS i	ζ_i	$1-\zeta_i$
Backhaul Resource Allocation Policy for link k	$Z(k)$	$1-Z(k)$
Traffic arrival rate (flows/sec) at location x	$\lambda^D(x)$	$\lambda^U(x)$
Max. rate of BS i BS at location x	$c_i^D(x)$	$c_i^U(x)$
Load density of BS i at location x	$\rho_i^D(x)$	$\rho_i^U(x)$
BS i max rate requirement for backhaul	$\tilde{c}_i^D$	$\tilde{c}_i^U$
Normalized load of BS i ($\zeta_i \to 1$ and $\zeta_i \to 0$)	ρ_i^D	ρ_i^U
Load of BS i	ρ_i^D/ζ_i	$\rho_i^U/(1-\zeta_i)$
Association chance of location x with BS i	$p_i^D(x)$	$p_i^U(x)$
Penalty indicator for congestion at BH link k	$\mathcal{J}^D(k)$	$\mathcal{J}^U(k)$
Penalty indicator for cross interf. between BS i,j	$\mathcal{I}_{ij}$	

process with arrival rate per unit area $\lambda(x)$[1]. Each new arriving request is for a *downlink (DL)* flow, with probability z^D, or *uplink* (UL) flow with probability $z^U = 1 - z^D$. Using a Poisson splitting argument [18], it follows that the above gives rise to 2 independent, Poisson flow arrival processes with rates

$$\lambda^D(x) = z^D \cdot \lambda(x), \quad \lambda^U(x) = z^U \cdot \lambda(x). \tag{1}$$

(A.2 - Flow characteristics) *Flow-sizes* (in bits) are drawn from a generic distribution with mean $1/\mu^D(x)$.

2.2 Access Network

(B.1 - Access network topology) We assume an area $\mathcal{L} \subset \mathbb{R}^2$ served by a set of base stations $\mathcal{B}$, that are either macro BSs (eNBs) or small cells (SCs).

(B.2 - Access Resource Allocation Policy) Each BS $i \in \mathcal{B}$ is associated with a total bandwidth w_i, and a resource allocation parameter $0 < \zeta_i < 1$ which reflects the amount of radio resources (e.g., time, frequency, space) available for DL transmissions. Without loss of generality, we focus on time resources, as e.g. in the context of the envisioned flexible TDD standard. [2] Hence, the (long-run) resources of BS i allocated to DL are $\zeta_i \cdot w_i$, whereas the UL ones are $(1-\zeta_i) \cdot w_i$, where ζ_i is a key *control variable* of our problem.

(B.3 - DL physical data rate) Each BS $i \in \mathcal{B}$ is associated with a transmit power P_i. It can deliver a *maximum* physical data transmission rate of $c_i^D(x, \zeta_i)$

[1] As we are interested in the aggregation of all flows from all locations x assosicated to BS i, even if flow arrivals at each location are not Poisson the Palm-Khintchine theorem [18] suggests that Poisson assumption could be a good approximation for the input traffic to a BS.

[2] Although traditional LTE systems only allow some fixed and predefined values for ζ_i (depending on the TDD configuration), we relax them to be more generally applicable.

to a user at location x in absence of any other flows served, given by Shannon capacity[3]

$$c_i^D(x, \zeta_i) = \zeta_i \cdot w_i \cdot \log_2(1 + \text{SINR}_i(x)), \tag{2}$$

where $\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j \neq i} G_j(x)P_j + N_0}$. N_0 is the noise power, and $G_i(x)$ represents the path loss and shadowing effects between the i-th BS and the UE located at x (as well as antenna and coding gains, etc.)[4]. We assume that effects of fast fading are filtered out, and that the total intercell interference at location x is static, and considered as another noise source, as in most aforementioned works [1, 15, 17].

(B.4 - Load density) We introduce the *load density* at x

$$\rho_i^D(x, \zeta_i) = \frac{\lambda^D(x)}{\mu^D(x) c_i^D(x, \zeta_i)}, \tag{3}$$

which is the contribution of location x to the total load of a BS i, when location x is associated with BS i.

(B.5 - BS load) Each location x is associated with routing probabilities $p_i^D(x) \in [0, 1]$, which are the probabilities that flows generated from a user at x get associated with (i.e., are served by) BS i. The effective load for BS i would be

$$\rho_i{}^D(\zeta_i) = \int_{\mathcal{L}} p_i^D(x) \rho_i^D(x, \zeta_i) dx. \tag{4}$$

Clearly, the BS loads ρ_i depend on (and are *coupled by*) the *new* control variables ζ_i, related to the UL/DL allocation problem. To make this relation explicit, in the following we will use the *normalized* load variables $\rho_i^D = \rho_i^D(\zeta_i = 1)$, i.e. the load when all resources are used for DL (similarly for UL). We are interested in the flow-level dynamics of this system, and model the service of DL flows at each BS as a queueing system with effective load (or *utilization*) $\frac{\rho_i^D}{\zeta_i}$.

(B.6 - Scheduling) Proportionally fair scheduling is often implemented in LTE networks due to its good fairness and spectral efficiency properties [19]. This can be modeled as an M/G/1 multi-class processor sharing (PS) system [18]. It is multi-class because each flow might get different rates for similarly allocated resources, due to different channel quality and modulation and coding scheme (MCS) observed at x.

(B.7 - Performance impact of BS load) The stationary number of flows in BS i is equal to $E[N_i] = \frac{\rho_i^D/\zeta_i}{1-\rho_i^D/\zeta_i}$ [18]. Hence, minimizing ρ_i^D/ζ_i minimizes $E[N_i]$, and by Little's law it also minimizes the per-flow delay for that BS [18]. Also, the throughput for a flow at location x is $\zeta_i \cdot c_i^D(x) \cdot (1 - \rho_i^D/\zeta_i)$. This observation is important to understand how the user's physical data rate $\zeta_i \cdot c_i^D(x)$ (related to

[3]We use Shannon capacity for clarity of presentation. However, our approach could be easily adapted to include modulation and coding schemes.

[4]In the UL, we assume that the Tx power of each user is P^{UE}, and slightly abuse notation for SINR, G, etc., as these don't play a major role later.

users at location x only) and the BS load ρ_i^D/ζ_i (related to *all* users associated with BS i) affect the optimal association rule (e.g., in Eq. (14)).

(B.8 - UL/DL association split) In the following, we will assume that a UE is able to associate with up to two BSs, one for its DL and one for UL traffic, as proposed in LTE Rel. 12 [20]. However, our framework is backward compatible when joint UL/DL association is required (see Section 5).

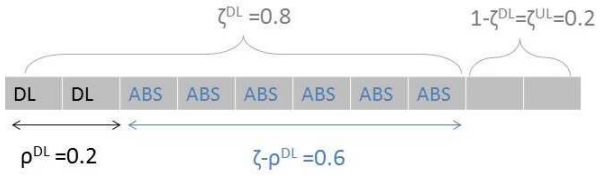


Figure 1: A frame example for a certain BS.

(B.9 - UL/DL cross interference avoidance) Without loss of generality, we assume that each BS i cross interferes with a subset of other BSs $\mathcal{C}_i \subseteq \mathcal{B} \setminus \{i\}$. In practice, a distance based rule, or alternatively the cell cluster concept, can be used to determine these sets. If i is on the DL and a BS $j \in \mathcal{C}_i$ on the UL (or vice versa) then these BSs might cause severe interference to each other (that invalidates assumption B.3). We refer to this as *cross interference*. A sufficient condition to avoid cross-interference is

$$\rho_i^D + \rho_j^U \leq 1, \forall i \in \mathcal{B}, j \in \mathcal{C}_i. \tag{5}$$

We explain the above condition here. Consider two such BSs i and j. If $\zeta_i = \zeta_j$ then there is no cross-interference, because i and j can synchronize their DL (and UL) slots to avoid it. If $\zeta_i \neq \zeta_j$, cross-interference might occur, but *it also depends on the effective loads*. ζ_i slots are *at most* used for DL. But out of these only $\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i = \rho_i^D$ will be busy (since $\frac{\rho_i^D}{\zeta_i}$ is the utilization of the downlink resources, according to B.5-B.7). The rest of the DL slots $(1 - \frac{\rho_i^D}{\zeta_i}) \cdot \zeta_i = \zeta_i - \rho_i^D$ could be blanked with ABS frames (see also Fig. 1). Similarly, the percentage of slots that j will be *active* on the UL is $\frac{\rho_j^U}{1-\zeta_j} \cdot (1 - \zeta_j) = \rho_j^U$ slots. Hence, if $\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i + \frac{\rho_j^U}{1-\zeta_j} \cdot (1 - \zeta_j) \leq 1$, there are enough different slots in a frame to schedule all DL and UL of i and j without any overlap. Taking care for all such links on the interference graph, gives us Eq.(5). Finally, we stress that this constraint applies to the long-term allocation policy of resources. The actual MAC scheduling may still allocate resources in those time slots to transmissions that are non-interfering.

2.3 Backhaul Network

(C.1 - Backhaul network topology) Each access network node (either eNB or SC) is connected to the core network through an eNB aggregation gateway via

a certain number of backhaul links that constitute the backhaul network. This connection can be either direct ("star" topology) or through one or more SC aggregation gateways ("mesh" topology).

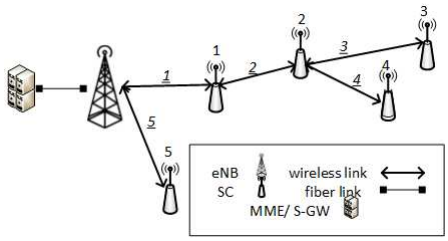


Figure 2: Future Backhaul topology of a HetNet.

Without loss of generality, we assume that there is a fiber link from the eNB to the core network, and focus on the set of capacity-limited backhaul links (wired or wireless) connecting SCs to the eNB, denoted as $\mathcal{B}_h$. We denote as routing path $\mathcal{B}_h(i)$ the set of all backhaul links $j \in \mathcal{B}_h$ along which traffic is routed from BS i to an eNB aggregation point, and we assume that it is *given* (e.g., calculated in practice as a Layer 2 (L2) spanning tree). For example, in Fig. 2, $\mathcal{B}_h(1) = \{1\}$, and $\mathcal{B}_h(3) = \{1, 2, 3\}$. We further denote as $\mathcal{B}(j)$ the set of all BS $i \in \mathcal{B}$ whose traffic is routed over backhaul link j. E.g., $\mathcal{B}(1) = \{1, 2, 3, 4\}$ and $\mathcal{B}(2) = \{2, 3, 4\}$ in Fig. 2.

(C.2 - Backhaul Resource Allocation Policy) Each $j \in \mathcal{B}_h$ backhaul link is associated with a total capacity $C_h(j)$. While traditional backhaul links are multiplexed using FDD, nowadays TDD gains more ground due to the performance improvements it promises [21]. So, in the context of TDD, we introduce the backhaul resource allocation parameter $0 < Z(j) < 1$, that splits the backhaul capacity of the j link between DL ($Z(j) \to 1$) and UL ($Z(j) \to 0$). Note that, backhaul links usually don't implement any particular scheduling algorithm, so they can be seen as a data "pipe".

(C.3 - Backhaul load) The DL load on a backhaul link j consists of the sum of DL loads of all BSs using that link ($i \in \mathcal{B}(j)$), divided by its offered backhaul capacity [15]

$$\sum_{i \in \mathcal{B}(j)} \frac{\frac{\rho_i^D}{\zeta_i} \cdot (\zeta_i \cdot \tilde{c}_i^D)}{Z(j) \cdot C_h(j)} = \sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D \cdot \tilde{c}_i^D}{Z(j) \cdot C_h(j)}. \tag{6}$$

where $\tilde{c}_i^D$ is a parameter use to "dimension" the BH link and corresponds to an estimate of the maximum DL total rate that BS i might request the backhaul to transport. A BS is characterized by its "peak" rate (often upper bounded by the maximum MCS available), and a "busy" rate when this BS serves many users [12]. The latter is usually quite smaller than the former, since users near the edge of the cell tend to bring the average rate down. However, the use of channel-based scheduling and related multi-user diversity gains suggest that conservatively setting

$\tilde{c}_i^D$ closer to its nominal peak value is safer. In practice, a BS can directly measure it.

(C.4 - Backhaul provisioning) Each BH link j is associated with a backhaul load (see C.3), that shall be maintained below 1 to prohibit backhaul congestion. As a result, each BH link is associated with a *backhaul constraint*:

$$\sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D \tilde{c}_i^D}{Z(j) \cdot C_h^D(j)} \prec 1, \ \forall j \in \mathcal{B}_h \tag{7}$$

Throughout this paper, we assume that the backhaul network is either *under-provisioned* if the capacity of *at least* one backhaul link is exceeded, or *over-provisioned* otherwise.

(C.5 - Interference-free Backhaul) Modern backhaul architectures are developed using (highly) directional P2P or P2MP static architectures [22]. These are planned topologies and thus cross interference between BH links with asymmetric UL/DL schedules can be considered negligible.

3 Joint Optimization

We start our discussion by ignoring the backhaul network (assuming it is provisioned), and attempt to solve the (i) *user association*, and (ii) *access resource allocation* problems, jointly. More specifically, we are interested in finding the optimal values for the variable ζ_i and $\rho_i^D, \rho_i^U, \forall i \in \mathcal{B}$. In Section 3.1 we define the feasible region of these variables, and then we introduce our objective function and the corresponding optimization problem. In Section 3.2 we sketch a convergent algorithm that decomposes it in smaller problems that can be efficiently tackled as shown in Section 3.3. In Section 3.4 we introduce and tackle the complete setting that also considers the (iii) *backhaul resource allocation* problem.

3.1 Feasible set, Objective and Optimization Problem 1.

The feasible region for our problem can be delimited by the requirement that the effective load of no BS being exceeded (see B.5).

Definition 1. ***(Feasible set)*** *If ϵ is an arbitrarily small positive constant, the feasible region of* $(\rho^D; \rho^U; \zeta) = ((\rho_1^D, \rho_2^D, \ldots, \rho_{\|\mathcal{B}\|}^D); (\rho_1^U, \rho_2^U, \ldots, \rho_{\|\mathcal{B}\|}^U); (\zeta_1, \zeta_2, \ldots, \zeta_{\|\mathcal{B}\|}))$

is

$$\mathcal{F} = \Big\{ (\rho^D, \rho^U, \zeta) \mid \rho_i^y = \int_{\mathcal{L}} p_i^y(x) \rho_i^y(x) dx, \tag{8a}$$

$$\sum_{i \in \mathcal{B}} p_i^y(x) = 1, \tag{8b}$$

$$0 \leq p_i^y(x) \leq 1, \ \ \forall x \in \mathcal{L}, \ \ y \in \{U, D\}, \tag{8c}$$

$$0 + \epsilon \leq \zeta_i \leq 1 - \epsilon, \tag{8d}$$

$$0 \leq \frac{\rho_i^D}{\zeta_i}, \frac{\rho_i^U}{1 - \zeta_i} \leq 1 - \epsilon, , \ \ \forall i \in \mathcal{B}, j \in \mathcal{C}_i \Big\} \tag{8e}$$

Lemma 3.1. *The feasible set $\mathcal{F}$ is convex.*

Proof. The proof for the feasible set $\mathcal{F}$ without the last two constraints can be found in [1]. Constraints (8d are linear, and constraint (8e) refers to the image of ρ under different perspectives. So they preserve convexity [23], and the complete feasible set remains convex. □

Following [1, 3] we extend the proposed objective that only considers the BS loads ρ_i, to also include the resource allocation variables $\zeta_i, \forall i \in \mathcal{B}$ (see B.2). The operator may weigh the importance of DL and UL traffic performance with a parameter $\tau \in [0, 1]$. α^D controls the amount of load balancing desired in the DL resources, and α^U in the UL. Let $\alpha = [\alpha^D; \alpha^U]$, where α^D and α^U can have different values.

Definition 2. ***(Objective function)*** *Our objective is*

$$\phi_\alpha(\rho, \zeta) = \sum_{i \in \mathcal{B}} \tau \frac{(1 - \frac{\rho_i^D}{\zeta_i})^{1-\alpha^D}}{\alpha^D - 1} + (1 - \tau) \frac{(1 - \frac{\rho_i^U}{1-\zeta_i})^{1-\alpha^U}}{\alpha^U - 1}, \textit{if } \alpha^D, \alpha^U \neq 1. \tag{9}$$

If α^D is equal to 1, the respective fraction must be replaced with $\log(1 - \frac{\rho_i^D}{\zeta_i})^{-1}$. The respective α-fair functions can capture different objectives such as maximizing spectral efficiency ($\alpha = 0$), throughput ($\alpha = 1$), mean per flow delay ($\alpha = 2$), and maxmin load-balancing ($\alpha \to \infty$); similarly for the UL.

This function, unlike the original one and some recent variants [3, 15, 17] is not convex, and thus the standard fixed point method or other convex solvers cannot be directly applied. However, the following lemma reveals a "hidden convexity" that can be exploited with decomposition methods.

Lemma 3.2. *The objective function $\phi_\alpha(\rho, \zeta)$ is a biconvex function, i.e., it is convex in ρ for fixed ζ, and versa.*

Proof. The objective function is the sum of the basic α function $\frac{(1-\frac{\rho}{\zeta})^{1-\alpha}}{\alpha - 1}$ over different BSs, with $(\rho, \zeta) \in \mathcal{F}$. When ζ is fixed this is the simplest form of the well

known α-fair function which is clearly convex in ρ. And so is the corresponding sum over all BSs (sum preserves convexity). For fixed ρ, the basic α function is also convex in ζ (it has non-negative second derivative, namely $2\rho\zeta^{-3}(1-\rho/\zeta)^{-\alpha}+\alpha\rho^2\zeta^{-4}(1-\rho/\zeta)^{-\alpha-1}\geq 0$), and so does its sum. □

Definition 3. ***(Optimization Problem 1)*** *The joint user association and radio resource allocation problem can be expressed*

$$\begin{aligned}&\min_{\rho,\zeta}\{\phi_\alpha(\rho,\zeta)|(\rho,\zeta)\in\mathcal{F}\},\\&\text{subject to Eq. }(5).\end{aligned}\tag{10}$$

Lemma 3.3. *Problem 1 is a biconvex minimization problem.*

Proof. This is a biconvex optimization problem since the objective function is biconvex on the (bi)convex feasible set $\mathcal{F}$, and the constraints are affine functions. □

3.2 Decomposition Algorithm for Optimization Problem 1.

Our nonconvex objective is block separable in $\rho^D,\ \rho^U$. Indeed, if we fix ζ, the problem decomposes in two simpler problems with variables ρ^D and ρ^U, that are coupled from constraint (5), and so we call ζ the *complicating* variable. Therefore, it makes sense to decompose the objective into two levels of optimization, following the *primal decomposition method* [24]. Specifically, at the lower level there are *two subproblems* that run in parallel, that aim to find the optimal values of ρ^{*D} and ρ^{*U}, namely $\rho=[\rho^{*D};\rho^{*U}]$, upon a fixed ζ. At the higher level we encounter the *master problem*, where we attempt to update (and eventually optimize), the complicating variable ζ. Note that constraint (5) only depends on ρ and thus does not affect the master problem. Formally, the subproblems and the master problem are

$$\min_{\rho}\{\phi_\alpha(\rho,\zeta)\}\ \text{ subj. to Eq.(5)}\qquad\text{(sub-problems)}\tag{11}$$

$$\min_{\zeta}\{\phi_\alpha(\rho,\zeta)\}\qquad\text{(master problem)}\tag{12}$$

The above decomposed problems are convex since Problem 1 is biconvex (see Lemma 3.4). Thus, they can efficiently be tackled through convex optimizers.

Our proposed iterative algorithm is sketched in Alg. 1. Convergence and stability are guaranteed if the two subproblems are solved on a faster timescale than the higher level master problem, so that at each iteration of a master problem both subproblems at a lower level have already converged [24]. In Section 3.3.1 we show how one can derive the optimal values ρ^*, whereas in Section 3.3.2 the sequence $\zeta^{(k)}$.

Lemma 3.4. *Algorithm 1 converges to the global optimal point of Problem 1.*

Algorithm 1 Decomposition Sketch of Problem 1.

1: **Repeat** until $\|\zeta^{(k)} - \zeta^{(k-1)}\| < \epsilon$.
2: *Update the master problem (Section (3.3.2)).*
3: Resource allocation: $\zeta \rightarrow$ DL, $1 - \zeta \rightarrow$ UL.
4: *Solve the two subproblems (Section (3.3.1)).*
5: Derive ρ^{*D} given the available resources (ζ).
6: Derive ρ^{*U} given the available resources ($1 - \zeta$).

Proof. Our proposed decomposition algorithm falls into the category of Alternate Convex Search (ACS) [25, 26], that is a special case of the popular Block Coordinate Decent (BCD) method [27]. There, starting from an initial feasible point, one attempts to minimize the objective by cyclically iterating through the different optimization directions with respect to one coordinate direction at a time. Precisely, in our case at the end of the k iteration it is

$$\phi_\alpha(\rho, \zeta^{(k)}) \prec \phi_\alpha(\rho, \zeta^{(k-1)}).$$

This will continue until convergence to a stationary point, where the gradient vanishes and the above inequality approaches equality. ACS algorithms in its simplest form suggest that the stationary point could be a saddle point, a local or global optimal [25]. However, Alg. 1 guarantees convergence to the global optimum due to the following two points.

(1) *Uniqueness of optimum point:* Optimization Problem 1 can be converted to a geometric programming (GP) problem, since both its objective and constraints can be written as a sum of posynomials terms composed of positive monomials, according to the transformation in [28]. Such problems have a single optimum. (The GP equivalent form of our problem is not convenient for decomposition, so we use this argument only to prove uniqueness, but not to solve the joint problem.)

(2) *Saddle point escape:* Our proposed algorithm can escape from potential saddle points, as discussed in Section 3.3.2. □

3.3 Subproblems and Master Problem.

3.3.1 Subproblem Optimization (Eq. (11))

We present here the DL subproblem only. The UL problem is symmetric. An efficient way to tackle the coupling constraints in a distributed implementation setup is to directly include the constraints in the objective as *penalty functions* that increase the objective when a cross-interference constraint is violated [23]. We can then solve the new *unconstrained* problem

$$\min_{\rho}\{\Phi(\rho, \zeta) = \phi_\alpha(\rho, \zeta) + \gamma \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^D + \rho_j^U - 1)^2\}, \tag{13}$$

where $\mathcal{I}_{ij}$ is the indicator variable that reveals whether BS i cross interferes with BS j ($\mathcal{I}_{ij}$=1, when $\rho_i^D + \rho_j^U > 1$) or not ($\mathcal{I}_{ij}$=0, otherwise) (see also B.9).

Quadratic penalty functions like the above are common [29] and preserve the convexity[5]. Parameter γ can be chosen as a large constant, introducing a "soft" constraint (i.e., cross-interference could be slightly exceeded, if this really improves our main objective), or be increased progressively, so as to converge to a "hard" constraint [29].

Theorem 3.5. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ denotes the optimal load vector, the optimal DL association rule for location x is*

$$i^D(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i^D(x)}_{\text{user knowledge}} \cdot \overbrace{P_i^D}^{\text{BS broadcast message}} \right) \tag{14}$$

where $P_i^D = \dfrac{\zeta_i \cdot \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}}{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j \in \mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}$.

Starting within a feasible point ρ and using increasing values for γ, these rules can be iteratively applied and will eventually converge to the optimal point ρ^*.

Proof. Problem (13) is convex. Let ρ^* be its optimal solution. A sufficient condition for optimality is if $\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle \geq 0$ for all $\rho \in \mathcal{F}$, where $\Delta\rho^* = \rho - \rho^*$. To write the remaining of the proof compactly with respect to the coupling constraints, we denote (only within the proof) $\zeta^D = \zeta, \zeta^U = 1 - \zeta$, $I(D) = \mathcal{I}_{ij}, I(U) = \mathcal{I}_{ji}$ and assume that L is either D or U ($L \in \{D, U\}$) with complementary value $\bar{L}$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic DL and UL cell coverage generated by (14) the respective optimal rules are $p_i^{*L}(x) = \mathbf{1}\left\{ i = i^L(x) \right\}$.

[5]This can be easily seen, since the function $(x + y - 1)^2$ has Hessian matrix the $[2, 2; 2, 2]$, and so it is positive semidefinite and convex.

Then, the inner product $\langle \nabla \phi \left(\rho^* \right), \Delta \rho^* \rangle$ is equal to

$$\sum_L \sum_{i \in \mathcal{B}} \left(\frac{1}{\zeta_i^L \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L} \right)^{\alpha^L}} + 2\gamma \sum_{j \in \mathcal{C}_i} I(L)(\rho_i^{*L} + \rho_j^{*\tilde{L}} - 1) \right) (\rho_i^L - \rho_i^{*L}) =$$

$$\sum_L \sum_{i \in \mathcal{B}} \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L} \right)^{\alpha^L} \sum\limits_{j \in \mathcal{C}_i} I(L)(\rho_i^{*L} + \rho_j^{*\tilde{L}} - 1)}{\zeta_i^L \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L} \right)^{\alpha^L}} \right) \cdot$$

$$\cdot \int_{\mathcal{L}} \rho_i^L(x) \left(p_i^L(x) - p_i^{*L}(x) \right) dx =$$

$$= \sum_L \int_L \frac{\lambda^L(x)}{\mu^L(x)} \sum_{i \in \mathcal{B}} \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L} \right)^{\alpha^L} \sum\limits_{j \in \mathcal{C}_i} I(L)(\rho_i^{*L} + \rho_j^{*\tilde{L}} - 1)}{\zeta_i^L c_i^L(x) \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L} \right)^{\alpha^L}} \right) \cdot$$

$$\cdot \left(p_i^L(x) - p_i^{L*}(x) \right) dx.$$

Note that in the DL i.e. $L = D$ (similarly in UL)

$$\sum_{i \in \mathcal{B}} p_i^D(x) \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i} \right)^{\alpha^D} \sum\limits_{j \in \mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i c_i^D(x) \left(1 - \frac{\rho_i^{*D}}{\zeta_i} \right)^{\alpha^D}} \right) \geq$$

$$\sum_{i \in \mathcal{B}} p_i^{D*}(x) \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i} \right)^{\alpha^D} \sum\limits_{j \in \mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i c_i^D(x) \left(1 - \frac{\rho_i^{*D}}{\zeta_i} \right)^{\alpha^D}} \right)$$

holds because $p_i^{*D}(x)$ is an indicator for the minimizer of $\frac{1+2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i} \right)^{\alpha^L} \sum\limits_{j \in \mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i c_i^L(x) \left(1 - \frac{\rho_i^{*D}}{\zeta_i} \right)^{\alpha^D}}$.

So, $\langle \nabla \Phi(\rho^*), \Delta \rho^* \rangle \geq 0$. □

Note that these rules are "device centric", i.e., they can be applied from the UE side in a distributed and iterative manner, as follows. At each iteration step *each BS broadcasts* the second term of the optimal rule (as indicated in Eq. 14), and each UE uses this message as well as its measured data rate to decide where to optimally associate, based on their product. (Similarly for the UL). Such broadcast quantities can be easily integrated through the newly proposed Access Network Discovery and Selection Function (ANDSF) mechanism or in the absolute/dedicated priority list mechanisms of LTE [30].

When the interference constraints for the BS i are not violated (i.e., $\mathcal{I}_{ij} = 0, \forall j \in \mathcal{C}_i$), the above rules state that the optimal downlink associations are the same as the one in [1]. However, when the BS i cross interferes with another BS, an additional term is added in the denominator that penalizes BS i making it less

preferable to users at location x. Note that the amount of penalization depends on the amount of *total* cross interference (sum term) from nearby BSs.

3.3.2 Master Pr. Update (Eq.(12))

Descent methods suggest:

$$\zeta^{(k+1)} = \zeta^{(k)} + t^{(k)} \Delta\zeta^{(k)}, \tag{15}$$

such that $\phi(\rho^*, \zeta^{(k+1)}) < \phi(\rho^*, \zeta^{(k)})$, where $\Delta\zeta^{(k)}$ is a *descent direction*, and $t^{(k)}$ a *step size*. The master step update for ζ could be performed centrally (e.g. at an SDN controller), or at each BS upon allowance for coordination with each other (e.g., exchanging ρ through the X2 interface, and/or using a distributed SDN controller environment) (see Section 5).

Nevertheless, since our objective is differentiable, we chose to apply the *Newton method* that provides the steepest descent direction in local Hessian norm, in order to speed up convergence. We also apply *backtracking line search* that determines the maximum amount to move along the search direction [23]. Finally, when stationarity is reached, we ensure that this is not a saddle point through a "noisy" gradient criterion: a noise vector with mean 0 is added to the gradient direction of stationary points that provably pushes them away from saddle points [31]. Due to space limitations, we refer the interested reader to [32] for more details.

3.4 (Joint) Optimization Problem 2 for Underprovisioned BH.

Introducing backhaul constraints, and flexible UL/DL resource allocation on each backhaul link, leads to a set of additional coupling constraints (where the coupling is now due to more than one BS utilizing the same BH link). Nevertheless, such constraints can also be tackled with appropriate penalty functions, which again leads to a decomposition of the problem, but now including an addition level (corresponding to an UL/DL update step at each BH link). Due to space limitations, and as the analysis and algorithms are extensions following similar logic to the previous section, we only provide here a brief description and refer the interested reader to [32].

Definition 4. ***(Optimization Problem 2)*** *The joint user association, radio resource allocation, and backhaul resource allocation problem can be expressed as*

$$\min_{\rho,\zeta,Z} \{\phi_\alpha(\rho,\zeta)_(\rho,\zeta,Z) \in \mathcal{F}\}, \textit{subj. to Eq.(}5\textit{) and (}7\textit{)}. \tag{16}$$

This is a multi-convex optimization problem (generalization of biconvex).The algorithmic sketch is shown in Algorithm 2. Convergence and stability can again be guaranteed if at each iteration of (either) master problems all the lower level problems have already converged. The optimal user association rules stay same in

nature, i.e. $i^D(x) = \arg\max_{i\in\mathcal{B}}\{c_i^D(x)\cdot P_i^D\}$, but now the BS broadcast message part P_i^D also includes the backhaul link penalties and is equal to:

$$\frac{\zeta_i\cdot\left(1-\frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}}{1+2\gamma\sum\limits_{k\in\mathcal{B}_h(i)}\tilde{c}_i^D\frac{\mathcal{J}^D(k)}{Z(k)\cdot C_h(k)}\left(\frac{\sum\limits_{l\in\mathcal{B}(k)}\rho_l^{*D}\tilde{c}_l^D}{Z(k)\cdot C_h(k)}-1\right)+\sum\limits_{l\in\mathcal{C}_i}\mathcal{I}_{ij}\cdot(\rho_i^{*D}+\rho_j^{*U}-1)}, \quad (17)$$

where $\mathcal{J}^D(k)$ indicates whether the k backhaul link is congested in the DL ($\mathcal{J}^D(k) = 1$ when $\frac{\sum_{i\in\mathcal{B}(k)}\rho_i\tilde{c}_i}{Z(k)C_h(k)} > 1$).

Algorithm 2 Decomposition Sketch of Problem 2.

1: **Repeat** until $\|Z^{(l)} - Z^{(l-1)}\| < \epsilon$.
2: *Update the master problem (Z).*
3: **Repeat** until $\|\zeta^{(k)} - \zeta^{(k-1)}\| < \epsilon$.
4: *Update the secondary master problem (ζ).*
5: *Solve the two subproblems (ρ).*

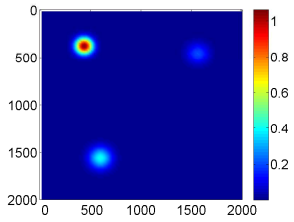
4 Simulations

In this section, we evaluate our proposed algorithms on example scenarios, and discuss related insights. We first consider a simple scenario with one macro BS and three SCs, in order to better elucidate the qualitative behavior of our algorithm, compared to standard practices, as well as better trace its performance benefits and where these come from. We then consider a larger network scenario and demonstrate that similar benefits can be observed there as well.

Scenario 1: We consider a $2\times 2\ km^2$ area. Fig.3 shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ (*flows/hour* per unit area) with 3 hotspots (blue implying low traffic and red high). We assume that this area is covered by three SCs (referred with BS numbers 1-3), and one macro cell (BS number 4). Without loss of generality, we assume that each SC offloads its traffic through a dedicated backhaul link (corresponding BH link numbers 1-3) to the macro BS, and that the macro BS cross interferes with all SCs (i.e., $C_4 = \{1,2,3\}$, $C_1 = C_2 = C_3 = \{4\}$, see B.9). We consider standard parameters as adopted in 3GPP [33], listed in Table II[6]. We set $\alpha^D = \alpha^U = 1$ to optimize user throughput. (We have also considered other values, with similar conclusions.)

Coverage Snapshots: We first look at the coverage maps that different schemes create. Figure 4(a), 4(b) depict the optimal user associations for fixed LTE-TDD configuration 1 that assumes static UL/DL timeslot ratio $4:4$ i.e., fixed $\zeta_i = 0.5, \forall i\in\mathcal{B}$. Similarly for the BH links $Z(j) = 0.5, \forall j\in\mathcal{B}_h$. As a first note,

[6]As for the sizes and ratios of different flows, as well as BH capacities, we can use different values in order to capture different simulation scenarios.



Variable	Value
$P_{eNB,SC,UE}$	43,24,12 dBm
$w_i,\ \forall i \in \mathcal{B}$	10 MHz
$C_h(j),\ \forall j \in \mathcal{B}_h$	100 MBps
N_0	-174 dBm/Hz
$1/\mu^{D,U}(x)$	100,25Kbytes
z^D, z^U	0.7,0.3
α^D, α^U	1,1

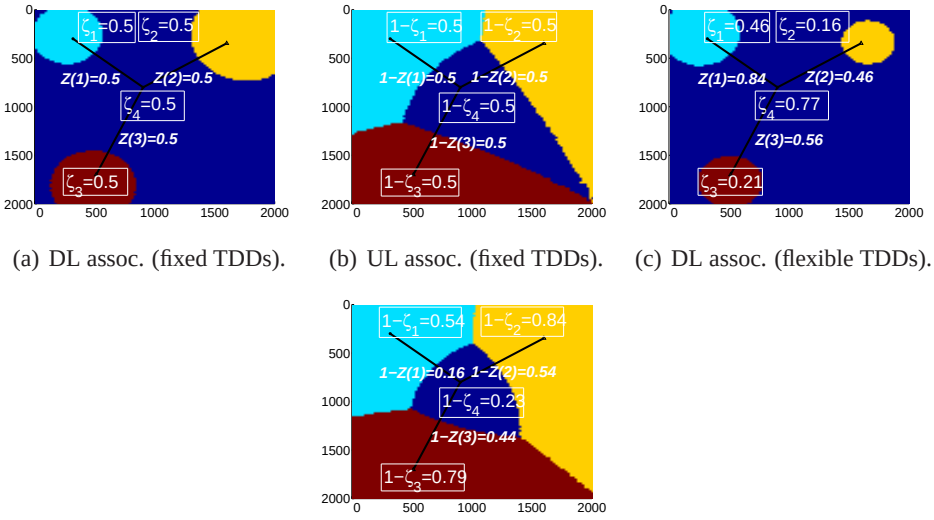
Figure 3 & Table 2: Traffic arrival rate and other simulation parameters.

we see that in DL most users are associated with the macro BS, and a few to SCs (macro BS attracts more DL users due to the higher transmit power). In the UL, users tend to form Voronoi cells (to minimize path loss and improve UL SINR). Secondly, we note that the DL coverage areas of the various SCs are decreased according to the corresponding traffic arrival intensity: e.g. SC 1 that serves the most intense hotspot (see Fig.3) has the smallest coverage area, while SC 3 which sees lower traffic intensity has the largest). The main reason is that the SCs have limited DL backhaul capacities that force some users to the far away macro BS. This alleviates the backhaul link congestion but hurts overall performance. At the same time, a high amount of the pre-configured UL backhaul resources might remain wasted (due, to assymetry in DL/UL traffic intensity for example).

Summarizing, the observed coverage maps for this scenario demonstrate two possible shortcomings of pre-configured TDD: (a) asymmetry in the DL/UL coverage areas and corresponding transmit powers suggest that a TDD allocation other than 50-50% could improve performance; (b) some (usually DL) user associations could be suboptimal, dictated by backhaul capacity limitations arising from the preconfigured fixed allocation on the BH, even if the total BH resources would suffice for the sum of both UL and DL traffic.

To explore these possibilities, we now relax the allocation variables ζ and Z (see B.2 and C.2) and apply our proposed algorithm. Clearly, in this simple example, a single-step improvement in either direction described above ((a) or (b)) could improve performance. We remind the reader that our proposed algorithm goes beyond this single step, alternating between optimizing coverage maps and TDD resource allocation, until it finds the best possible combination. The resulting coverage maps (i.e. optimal ρ values) and radio/BH allocations (optimal ζ and Z values) are shown in Fig. 4(c), 4(d). We first note that macro BS increases its $\zeta_4 = 0.77$ to serve more DL users, and SC increase their UL resources $1-\zeta_1 = 0.54, 1-\zeta_2 = 0.84, 1-\zeta_3 = 0.79$ to serve more UL, bewaring to avoid *cross interference*. Interestingly, such an allocation simultaneously improves both UL and DL performances (we will explicitly show this later). Also, the DL BH allocated resources ($Z(j)$) are increased to accommodate more DL traffic, while ensuring not to exceed a maximum value that would congest the UL.

User-centric performance: We now go beyond the above qualitative behavior



(a) DL assoc. (fixed TDDs). (b) UL assoc. (fixed TDDs). (c) DL assoc. (flexible TDDs).

(d) UL assoc. (flexible TDDs).

Figure 4: DL and UL user associations for different scenarios ($\tau = 0.5$).

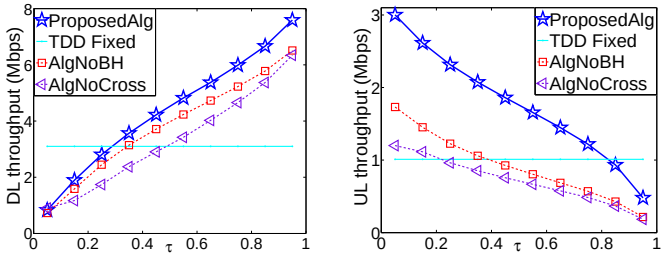
and evaluate the quantitative benefits. We first focus on user-centric performance and consider various τ values (we remind the reader that τ is a parameter that balances the importance of DL vs UL performance). We compare the performance of the following main schemes. (*ProposedAlg*): our proposed algorithm; (*TDD Fixed*): the optimal allocation algorithm of [15] with equal, pre-confiigured UL/DL resources on both radio access and BH. To better understand the importance of considering the cross-interference and BH capacity constraints, we also include results for the following schemes. (*AlgNoCross*): jointly optimal allocation, but not taking cross-interference into account. If there is an eventual asymmetry in the optimal UL/DL schedules, potential cross-interference is included in the SINR to capture its impact. (*AlgNoBH*): jointly optimal allocation without considering the backhaul constraints. Here, we assume that all BSs associated with a BH link that is congested decrease their performance proportionally to the amount of congestion.

In Fig 5 we depict the DL and UL user throughput as a function of τ in different scenarios. It is easy to see that our *ProposedAlg significantly outperforms* the *TDD fixed* policy by up to $2.5 - 3\times$. What is more, for most intermediate τ values, it is able to simultaneously improve both DL and UL performance. As τ increases further, the emphasis of *ProposedAlg* moves exclusively to the DL (and vice versa) which is consistent with our expectations, unlike the fixed TDD scheme where DL and UL performances are optimized independently of τ (decoupled objective).

Regarding the impact of the cross interference constraint, *AlgNoCross* can still offer some improvement on the DL for $\tau > 0.5$, compared to the baseline (*TDD Fixed*). However, it does so with a significant penalty on UL performance (up to $3\times$ worse), which is the most sensitive to cross-interference (this DL-to-UL inter-

ference is a key problem for future Flexible TDD [34]). This underlines the importance of directly considering cross interference constraints in our optimization framework through Eq.(5). Finally, the performance of *AlgNoBH* shows similar behavior, where it can sometimes provide better performance for the DL or the UL (compared to *TDD fixed*) but not both.

Summarizing, the following important conclusions can be drawn from the above analysis: (a) jointly optimal allocation of user association and DL/UL radio resources can actually lead to considerable performance degradation, unless cross-interference is taken explicitly into account; (b) a jointly optimal allocation, even with cross-interference taken into account, might still be quite suboptimal, if the DL/UL resources on the BH are not also optimized to conform to the new load requirements imposed by the BSs; (c) joint optimization of all these dimensions is feasible, and can offer significant performance improvement for both DL/UL.



(a) DL throughput. (b) UL throughput.

Figure 5: User-centric Performance.

Network-centric performance. Table 3 considers the performance improvements in the same comparison scenario (*ProposedAlg* and *TDD Fixed* [15]), but now from the network perspective when $\tau = 0.5$. We consider two metrics: Spectral Efficiency (SE) in terms of bits/s/Hz, and Load Balancing (LB) in terms of mean square error between different BS loads, similar to what is assumed in [15]. DL/UL spectral efficiency improve up to 44% since *flexible TDD better allocates the resources* with respect to the heterogeneous transmit powers that help physical data rates improve (see B.2-B.3). It also considers related traffic statistics and asymmetries across users (see A.1-A.2) by diminishing the BS load fluctuations (e.g., BS under/over utilizations) and thus LB is improved. It is interesting to note that simultaneous improvement of these metrics implies improvement in user performance, as showed previously and explained in B.7.

Table 3: Network (SE,LB) Performance ($\tau = 0.5$)

	Downlink		Uplink	
Performance.	SE	LB	SE	LB
Percentage $\%$ of improvement.	42	16	44	54

Scenario 2: Having highlighted the different tradeoffs and sources of performance improvement in the basic scenario above, we now turn our attention to a larger network topology consisting of 4 macro BSs and 13 SCs. Without loss of generality, we now consider uniform traffic demand. Considerable performance improvements can be observed in this scenario as well (e.g. 86% better UL user performance). Relative lower improvement values compared to the smaller Scenario 1 are mainly due to: (a) not all BSs experience bad performance now so even if *ProposedAlg* considerably improves the performance of the problematic BSs, average performance is not as affected; (b) the *additional cross interference constraints* posed from the neighboring clusters. Due to space limitations, we refer the interested reader to [32] for more details on this scenario.

Table 4: User (UE) and Network (SE, LB) Performance ($\tau = 0.5$)

	Downlink			Uplink		
Scenario.	UE	SE	LB	UE	SE	LB
Percentage $\%$ of improvement.	29	39	4	86	42	51

5 Discussion and Future Work

Decomposition order. While our proposed decomposition is not the only possible decomposition, we believe this lends itself to a natural implementation between different network elements. User association is proposed to run in the fastest timescale to adapt to the high traffic fluctuations across different locations and users. The load of a single BS depends on the sum of its attached users and is subject to fewer fluctuations. It only has to react to (slower) traffic shifts of the aggregate loads, by updating its ζ parameter accordingly. Finally, a backhaul link further aggregates the traffic of multiple BS, and can update its optimal allocation at an even slower timescale.

Scalability and Flexibility. Our user association rules are "device centric", i.e., the user is able to select where to associate *based on own measurements (e.g. SINR) and BS-transmitted information*. This is inline with user association in current LTE systems, where user association depends on device centric information (e.g. SINR measurements) but also BS-transmitted information (e.g. priority lists of BSs to monitor). These rules are: *scalable*, (constant amount of the BS broabcast messages irrespective of the number of users, backhaul topology, and cross-interference map), *simple* (constant complexity of the rule with respect to the number of BSs), and offer *flexible performance* (defined from α values).

Distributed SDN-based control. The two master algorithms, for access and backhaul TDD allocation, require by some centralized knowledge. While these can run at a slower time scale without jeopardizing the performance of the algorithm, a hierarchical or fully distributed implementation, based on hierarchical

SDN controllers could be envisioned [35]. A local SDN controller could, e.g., be responsible for a smaller "cluster" of MC, SCs, and their backhaul. The master problem could be further decomposed into further subproblems blocks, each solved by the respective SDN controller (and then aligned through communication with a main SDN controller). We intend to investigate such a scenario in future work.

Joint UL/ DL association: Our framework is also applicable when DL and UL traffic at a location x have to be offloaded to the same BSs (see B.8), by requiring $p_i^D(x) = p_i^U(x)$ in the association rule derivations (see [32] for the resulting rules). We defer to future work other similar splits, e.g., for control/data channels, or best effort/dedicated traffic [19].

6 Conclusion

In this paper, we formulated a novel algorithm that carefully studies the coupled problems of (i) user association, TDD (ii) access, and (iii) backhaul resource allocation under the emerging *backhaul* and *cross interference* constraints. Using optimization theory we proved that under certain circumstances it converges to the global optimum. Simulation results corroborate the correctness of our framework and reveal promising qualitative and quantitative results.

References

[1] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha -optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, 2012.

[2] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[3] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "An analytical framework for optimal downlink-uplink user association in hetnets with traffic differentiation," in *Proc. IEEE Globecom*, 2015.

[4] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc. IEEE Globecom*, 2014.

[5] V. Pauli and E. Seide, *Dynamic TDD for LTE-A and 5G*, 2015.

[6] G. 36.133, "Evolved universal terrestrial radio access (E-UTRA) and radio access network (E-UTRAN); overall descriptiom," 2012.

[7] R. Sivaraj, I. Broustis, N. K. Shankaranarayanan, V. Aggarwal, R. Jana, and P. Mohapatra, "A QoS-enabled holistic optimization framework for LTE-advanced heterogeneous networks," in *Proc. IEEE Infocom*, 2015.

[8] M. Ding, D. L. Perez, A. V. Vasilakos, and W. Chen, "Dynamic TDD transmissions in homogeneous small cell networks," in *Proc. IEEE ICC Communications Workshops*, 2014.

[9] H. Ji, Y. Kim, S. Choi, J. Cho, and J. Lee, "Dynamic resource adaptation in beyond LTE-A TDD heterogeneous networks," in *Proc. IEEE ICC Communications Workshops*, 2013.

[10] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic uplink-downlink optimization in TDD-based small cell networks," in *International Symposium on Wireless Communications Systems*, 2014.

[11] 3GPP, "TS 36.300, Release 13 (version 13.2.0)," 2016.

[12] *Backhaul technologies for small cells*, Small Cell Forum, 2014.

[13] D. Chen, T. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Transactions on Wireless Communications*, 2015.

[14] J. Ghimire and C. Rosenberg, "Revisiting scheduling in heterogeneous networks when the backhaul is limited," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2015.

[15] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Optimal downlink and uplink user association in Backhaul-limited HetNets," in *Proc. IEEE Infocom*, 2016.

[16] M. Shariat, E. Pateromichelakis, A. Quddus, and R. Tafazolli, "Joint TDD backhaul and access optimization in dense small cell networks," *IEEE Transactions on Vehicular Technology*, 2013.

[17] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE Journal on Selected Areas in Comm.*, 2011.

[18] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems*. Imperial college press, 2010.

[19] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communication Surveys and Tutorials*, 2013.

[20] 3GPP, "TR 36.842, Release 12 (version 12.0.0)," 2014.

[21] E. Metsala and J. Salmelin, *Mobile Backhaul*. Wiley, 2012.

[22] Cambridge Broadband Networks, "Solutions mobile backhaul," 2015.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[24] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, 2006.

[25] R. E. Wendell and A. P. Hurter, "Minimization of a non-separable objective function subject to disjoint constraints," *Journal on Operation Research*, 1976.

[26] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, 2007.

[27] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.

[28] C. A. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications*. Springer-Verlag New York, Inc., 2000.

[29] Z. G. Raphael T. Haftka, *Elements of Structural Optimization*. Springer Netherlands, 1992.

[30] S. Sesia, I. Toufik, and B. M., *LTE - The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*. Wiley, 2011.

[31] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points- online stochastic gradient for tensor decomposition," in *Proc. of the 28th Conference on Learning Theory*, 2015.

[32] "https://www.dropbox.com/s/dd1kljjv1qsa124/report.pdf?dl=0."

[33] 3GPP, "TR 36.931 Release 13 (version 13.2.0)," 2016.

[34] Z. Shen, A. Khoryaev, E. Eriksson, and X. Pan, "Dynamic uplink-downlink configuration and interference management in TD-LTE," *IEEE Communications Magazine*, 2012.

[35] K. Phemius, M. Bouet, and J. Leguay, "Disco: Distributed SDN controllers in a multi-domain environment," in *IEEE Network Operations and Management Symposium*, 2014.