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

Research Report 16-318

User Association in over- and under- provisioned Backhaul HetNets

April 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.

User Association in over- and under- provisioned Backhaul HetNets

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

Abstract

Operators, struggling to continuously add capacity and upgrade their architecture to keep up with data traffic increase, are turning their attention to denser deployments that improve spectral efficiency. Denser deployments make the problem of user association challenging, and much work has been devoted to finding algorithms that strike a tradeoff between user quality of service (QoS), and network-wide performance (load-balancing). Nevertheless, the majority of these algorithms typically consider simple setups with a single type of traffic, usually elastic non-GBR (Guaranteed Bit Rate). They also focus on the radio access part, ignoring the backhaul topology and potential capacity limitations. Backhaul constraints are emerging as a key performance bottleneck in future networks, partly due to the continuous improvement of the radio interface, and partly due to the need for inexpensive backhaul links to reduce capital and operational expenditures. To this end, we propose an analytical framework for user association that jointly considers radio access and backhaul network performance. Specifically, we derive an algorithm that takes into account spectral efficiency, base station load, backhaul link capacities and topology, and two traffic classes (GBR and non-GBR) in both the uplink and downlink directions. We prove analytically an optimal user association rule that ends up maximizing either an arithmetic or a weighted harmonic mean of the achieved performance along different dimensions (e.g. UL and DL performance or GBR and non-GBR performance). We then use extensive simulations to study the impact of (i) traffic differentiation, and (ii) backhaul capacity limitations and topology on key performance metrics.

Index Terms

hetnets; backhaul; optimization; traffic differentiation; user-association; load balancing; spectral efficiency.

Contents

List of Figures

1 Introduction

Driven by the exponential growth in wireless data traffic, operators are increasingly considering denser, heterogeneous network (HetNet) deployments. In a HetNet, a large number of small cells (SC) are deployed along with macrocells to improve spatial reuse [1–3]. The higher the deployment density, the better the chance that a user equipment (UE) can be associated with a nearby base station (BS) with high signal strength, and the more the options to balance the load. At the same time, denser deployments experience high spatio-temporal load variations, and require sophisticated user association algorithms. There are two key, often conflicting concerns when assigning UEs to a BS: (i) maximizing the spectral efficiency, and (ii) ensuring that the load across BSs is balanced to improve the utilization efficiency, and preempt congestion events. The former is usually achieved by associating the UE to the BS with maximum SINR: this association rule was the base up to LTE (Long-Term Evolution)-release 8. While this rule also maximizes the *instantaneous* rate of a user (i.e., the modulation and coding scheme - MCS - supported), it reflects user QoS only when the BS is lightly loaded. However, user performance, in terms of *per flow delay*, may be severely affected if the BS offering the best SINR is congested [4, 5].

As a result, a number of research works have studied the problem of user association in heterogeneous networks, optimizing user rates [6, 7], balancing BS loads [8], or pursuing a weighted tradeoff of them [9]. For instance, a distributed user-association algorithm is proposed in [10], where the global outage probability and the long term rate maximization are well studied, in the context of load balancing. The authors in [11] propose a framework that studies the interplay of user association and resource allocation in future HetNets, by formulating a non-convex optimization problem and deriving performance upper bounds. Range-expansion techniques, where the SINR of lightly loaded BSs is biased to make them more attractive to the users are also popular [2, 3]. Finally, a framework that has received much attention is [9]. This framework jointly considers a family of objective functions, each of which directs the optimal solution towards different goals (e.g. throughput optimal, delay-optimal, load balancing, etc.), using an iterative algorithm. [12–14] extend this framework to further include energy management, e.g., by switching off under-loaded BSs.

Nevertheless, the majority of these works are relatively simplified, not taking into account key features of future networks. Firstly, most existing studies only consider homogeneous traffic profiles. For example, [9, 12, 15] assume that all flows generated by a UE are “best-effort” (i.e. elastic). However, modern and future networks will have to deal with high traffic differentiation, with certain flows

This work was partially supported by the project “Network-level Optimizations for Small Cell Networks", funded by Intel Mobile Communications.

being able to require specific, *dedicated*[1] (i.e., non-elastic) resources [16]. Such dedicated flows do not share BS resources like best-effort ones, are subject to admission control, and sensitive to different performance metrics [17]. Secondly, the majority of related studies only consider downlink (DL) traffic. Uplink (UL) traffic is becoming important, due to symmetric (e.g. social networking) applications, Machine-Type Communication (MTC), etc. Yet, due to the asymmetric transmit powers of UEs and BSs, leading to different physical data rates, the BS which is optimal for DL traffic might lead to severely degraded performance for UL traffic. Summarizing, a proper user-association scheme should consider all the above dimensions, and attempt to strike an appropriate tradeoff between them.

On top of that, most related works focus on the radio access part (e.g., considering the user rate on the radio interface or BS load), ignoring the backhaul (BH) network. While this might be reasonable for legacy cellular networks, given that the macrocell backhaul is often over-provisioned (e.g., fiber), this might be quite suboptimal for future cellular networks. The considerably higher number of small cells, and related Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) suggest that backhaul links will mostly be inexpensive wired or wireless (in licensed or unlicensed bands), and underprovisioned [18]. Multiple BS might also have to share the capacity of a single backhaul link due to, e.g, point-to-multipoint (PMP) or multi-hop mesh topologies to the aggregation node(s) [19]. Finally, various BS-coordinated schemes have been proposed in the literature as a promising way to better use the available spectrum and further improve system performance, e.g., enhanced Inter-Cell Interference Coordination (eICIC) [20,21] and Coordinated Multi-Point (CoMP) transmission [22] scenarios. Such schemes are expected to further stress the backhaul network capacities. Hence, as the radio access technologies are constantly improving, it is argued that the backhaul network will emerge as a major performance bottleneck, and user association algorithms that ignore the backhaul load and topology can lead to poor performance [23].

As a result of this increasing focus on the backhaul, some recent works have appeared that attempt to jointly consider radio access and backhaul. These are mostly concerned with joint scheduling issues (for in-band or PMP backhaul links) [23, 24], signaling overhead and performance tradeoffs for cooperative multi-point communication [25], Software-Defined-Networking (SDN)-based implementation flexibility [26], or propose some simple heuristics to include the impact of the backhaul network on user association [27]. Finally, Chen et al. attempt to derive the total expected delay by taking into account retransmission over the wireless links, as well as the backhaul delay in the wireless backhaul links [28]. Nevertheless, to our best knowledge, none of these works formally addresses the problem of optimal user association in future and potentially backhaul-limited HetNets.

To this end, we revisit the user association problem, jointly considering the radio access and backhaul networks. Specifically, our main contributions can be

[1] In terms of LTE systems, dedicated flows are differentiated by their QoS class (QCI) ranging from 1 to 4, whereas best-effort from 5 to 9 [16].

summarized as follows

1) We use the popular framework of α-optimal user association [9] as our starting point, and considerably extend it to include (i) traffic differentiation, (ii) UL traffic, and (iii) backhaul topology and capacity constraints.

2) We then analytically prove different association rules, depending on whether UL and DL traffic of the same UE can be “split” to different BSs or not [29]. Interestingly, depending on this UL/DL “split" the derived rules end up maximizing either an arithmetic or a weighted harmonic mean of the optimal association rules per problem dimension.

3) We use our framework to investigate the various tradeoffs arising in this complex association problem, and provide some initial insights and guidelines about the impact of traffic differentiation and backhaul limitations in optimal user-association policies for future HetNets.

4) Our results also highlight some shortcomings of future HetNets, and indicate potential extensions to tackle them within our framework. These include the need for joint radio access and Layer 3 routing on the transport (backhaul) network, and dynamic allocation of access as well as backhaul resources (e.g., in the context of dynamic TDD).

The remainder of the paper is organized as follows: Section 2 describes the system model and related assumptions. In Sections 3 and 4 we derive the optimal user-association policies for provisioned and under-provisioned backhaul network. In Section 5 we simulate our proposed optimal association rules and attempt to shed some light on the impact of traffic differentiation, backhaul topology and capacity on system performance. Section 6 discusses potential extensions of our framework, and Section 7 concludes the paper.

2 System Model and Assumptions

In the following, we describe our traffic arrival model (Section 2.1), the discuss our assumptions related to the access (Section 2.2) and backhaul networks (Section 2.3).

We use a similar problem setup as the one used in a number of related works [9, 12, 13, 30], and extend it accordingly. To keep notation consistent, for all variables considered a first superscript “D" and “U" refers to downlink (DL) and uplink (UL) traffic, respectively. A second superscript “b” or “d” refers to best-effort and dedicated 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.

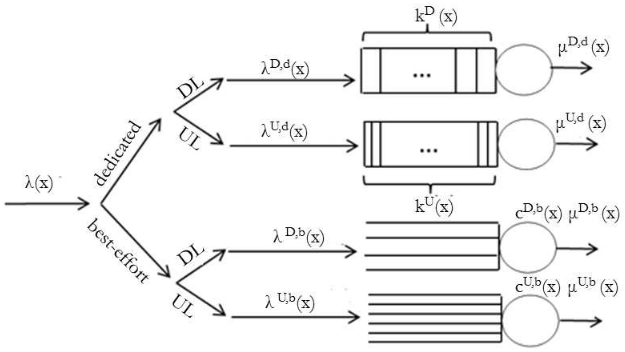


Figure 1: Access network queuing systems for different flows.

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 process with arrival rate per unit area $\lambda(x)$[2]. This inhomogeneity facilitates the creation of "hotspot" areas. Each new arriving request is for a *downlink (DL)* flow, with probability z^D, or *uplink (UL)* flow with probability $z^U = 1 - z^D$. Each DL (or UL) flow can furher be a *best-effort* flow (e.g., file download) with probability z^b, or *dedicated* flow (e.g., a VoIP call), with probability $z^d = 1 - z^b$. z^D and z^b are input parameters that depend on the traffic mix.

Using a Poisson splitting argument [31], it follows that the above gives rise to 4 independent, Poisson flow arrival processes with respective rates

$$\lambda^{D,b}(x) = z^D \cdot z^b \cdot \lambda(x), \quad \lambda^{D,d}(x) = z^D \cdot z^d \cdot \lambda(x) \tag{1}$$
$$\lambda^{U,b}(x) = z^U \cdot z^b \cdot \lambda(x), \quad \lambda^{U,d}(x) = z^U \cdot z^d \cdot \lambda(x), \tag{2}$$

($\lambda^{D,b}(x)$ for the downlink best-effort flows, $\lambda^{U,b}(x)$ for the uplink best-effort flows, etc.).

(A.2 - Best effort flow characteristics) Each *best-effort* flow is associated with a *flow-size* (in bits) drawn from a generic distribution with mean $1/\mu^{D,b}(x)$. This can model heterogeneous flow characteristics across locations.

(A.3 - Dedicated flow characteristics) Each *dedicated* flow has a *required data-rate* (in bits per second) that is drawn from a generic distribution with mean $B^D(x)$. This rate must be guaranteed by the network throughout the flow's duration. This duration (in seconds) is another, independent random variable with mean $1/\mu^{D,d}(x)$.

[2] Without loss of generality, we do not distinguish between users at location x, as we assume that all users/flows related to location x are treated similarly.

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). These together constitute the access network.

(B.2 - DL resources) Each BS $i \in \mathcal{B}$ is associated with a transmit power P_i and a total downlink bandwidth w_i^D. Out of the total bandwidth, $\zeta_i^D \cdot w_i^D$ is allocated to best-effort traffic and $(1-\zeta_i^D) \cdot w_i^D$ for dedicated traffic ($0 \leq \zeta_i^D \leq 1$). Throughout this paper, we will assume that this allocation is static, at least for a given time window of interest (based on long term traffic characteristics and operator policy). Dynamically updating the ζ_i^D parameters could further improve performance, but is related more to the MAC scheduler of each BS and is out of the scope of this paper. Nevertheless, in Section 6, we discuss how one could include this in our framework.

(B.3 - DL physical data rate) BS i can deliver a *maximum* physical data transmission rate of $c_i^{D,b}(x)$ to a user asking for a best-effort flow at location x, in absence of any other best-effort flows served, which is given by the Shannon capacity[3]

$$c_i^{D,b}(x) = \zeta_i^D \cdot w_i^D \cdot \log_2(1 + \text{SINR}_i(x)), \tag{3}$$

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. Our model assumes that the total intercell interference at location x is static, and considered as another noise source, as is previously considered in most aforementioned works [9, 12].

The next 4 points (B.4-B.7) describe the scheduling and performance model for best effort traffic only. We return to dedicated traffic in (B.8-B.9).

(B.4 - Best effort load density) We introduce the *load density* for best effort flows, at different locations x,

$$\rho_i^{D,b}(x) = \frac{\lambda^{D,b}(x)}{\mu^{D,b}(x) c_i^{D,b}(x)}, \tag{4}$$

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

(B.5 - Best effort load) Each location x is associated with routing probabilities $p_i^{D,b}(x) \in [0, 1]$, which are the probabilities that best effort DL flows generated for

[3] We use Shannon capacity for clarity of presentation. However, our approach could be easily adapted to include modulation and coding schemes (MCS). Furthermore, capacity improving technologies, e.g., the use of MIMO, and modifications to this capacity formula are othogonal to our framework.

[4] In the case of 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 in the remaining discussion.

users at location x get associated with (i.e., are served by) BS i. We can thus define the *total best effort load* $\rho_i^{D,b}$ for BS i as

$$\rho_i{}^{D,b} = \int_{\mathcal{L}} p_i^{D,b}(x)\rho_i^{D,b}(x)dx. \tag{5}$$

Similarly to [4, 9], we are interested in the *flow-level dynamics* of this system, and model the service of DL best-effort flows at each BS as a queueing system with load $\rho_i^{D,b}$ shown in Fig. 1. Finally, since we are interested in the aggregation of all flows at BS level (i.e., all flows from all locations x assosicated to BS i), even if flow arrivals at each location are not Poisson (as in A.1), the Palm-Khintchine theorem [31] suggests that Poisson assumption could be a good approximation for the input traffic to a BS.

(B.6 - Best effort scheduling) Proportionally fair scheduling is often implemented in 3G/4G networks for best-effort flows, due to its good fairness and spectral efficiency properties [16]. This can be modeled as an M/G/1 multi-class processor sharing (PS) system (see, e.g., [4,9,12]). It is multi-class, because each flow might get different rates for similarly allocated resources, due to different channel quality and MCS at x. Channel-based scheduling could also be included in the model and can be accounted for using a multiplicative factor in the average service rate [32].

(B.7 - Performance for best effort flows) The stationary number of flows in BS i is equal to $E[N_i] = \frac{\rho_i^{D,b}}{1-\rho_i^{D,b}}$ [31]. Hence, minimizing $\rho_i^{D,b}$ minimizes $E[N_i]$, and by Little's law it also minimizes the per-flow delay for that base station [31]. Also, the throughput for a flow at location x is $c_i^{D,b}(x)(1-\rho_i^{D,b})$. This observation is important to understand how the user's physical data rate $c_i^{D,b}(x)$ (related to users at location x only) and the BS load $\rho_i^{D,b}$ (related to *all* users associated with BS i) affect the optimal association rule.

(B.8 - Dedicated traffic load density) Unlike best-effort flows which are elastic, dedicated flows are subject to admission control, since they require some resources for exclusive usage in order to be accepted in the system. Specifically, let $c_i{}^{D,d}(x)$ denote the maximum offered rate to users at location x corresponding to dedicated flows only (referred to $(1-\zeta_i)$ - see B.3 above). If each flow at x demands, on average, a rate of $B^D(x)$ (see A.3), then at most $k_i^D(x) = \frac{c_i{}^{D,d}(x)}{B(x)}$ dedicated flows from x could be served in parallel by BS i (assuming again *no other flows in the system*), and any additional flows would be rejected[5]. Similarly to the best effort case (B.4), we can define a system load density for dedicated traffic at x

$$\rho_i{}^{D,d}(x) = \frac{\lambda^{D,d}(x)}{\mu^{D,d}(x)k_i^D(x)} = \frac{\lambda^{D,d}(x)\cdot B^D(x)}{\mu^{D,d}(x)\cdot c_i{}^{D,d}(x)}. \tag{6}$$

[5] In fact, since the rate requirement for each flow is a random variable, using its mean $B^D(x)$ in the denominator yields a lower bound for $k_i^D(x)$ (by Jensen's inequality), which can be used as a conservative estimate.

Hence, a different number of resources $k_i^D(x)$ can be offered to different locations x, depending on the rate demand $B^D(x)$ as well as the channel quality (rate ${c_i}^{D,d}(x)$) at location x.

(B.9 - Dedicated traffic performance) Given the above heterogeneous blocking model for dedicated flows, we can approximate the allocation of BS i dedicated resources with an M/G/k/k (or k-loss) system, where the total load $\rho_i^{D,d}$ can be calculated as in (B.5) and Eq. (5), using the density of Eq. (6) and corresponding routing probability $p_i^{D,d}(x)$ for dedicated flows (see also Fig. 1). It is known that for M/G/k/k systems, minimizing $\rho_i^{D,d}$ is equivalent to minimizing the blocking probability for new flows [31]. This observation is important to understand that a similar tradeoff (as in B.7) exists between choosing a BS at x that maximizes $k_i^D(x)$ (related only to flow and channel characteristics at x) and choosing a BS whose *total* load $\rho_i^{D,d}$ (related to *all* users attached to BS i).

(B.10 - UL/DL association split) We investigate two scenarios, depending on the whether a UE is allowed to be attached to different BSs for its DL and UL traffic [29]:

Split UL/DL: Each UE can be associated to different BSs for its DL and UL traffic. This allows one to optimize UL and DL performance independently [33].

Joint UL/DL: Each UE must be associated with the same BS for both UL and DL traffic. This is the standard practice in current networks.

2.3 Backhaul Network

(C.1 - Backhaul network topology) Each access network node (either eNB or SC) is connected to the core network through the 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 ("tree" topology). Fig. 2 shows such a backhaul routing topology.

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. 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. In the case of a star topology, there is exactly one (unique) backhaul link used for each BS (i.e., $\|\mathcal{B}_h(i)\| = \|\mathcal{B}(j)\| = 1, \forall i, j$). Finally, we assume that the backhaul route for each BS is *given*, e.g., calculated in practice as a Layer 2 (L2) spanning tree, and is an input to our problem. In Section 5, we highlight some limitations of L2 backhaul routing.

(C.2 - Backhaul load) Each backhaul link $j \in \mathcal{B}_h$ is characterized by a DL and UL capacity, denoted as $C_h^D(j)$ and $C_h^U(j)$ bps. The capacity on the UL and DL might be the same or different (e.g., Frequency-Division Duplex (FDD), or fixed/dynamic Time-Division Duplex (TDD) systems [34]). Backhaul links usually don't implement any particular scheduling algorithm, and can be seen as a data "pipe".

Without loss of generality, we focus on a scenario with only best-effort traffic. This not only keeps our backhaul model tractable as we shall see later, but also allows us to better understand the impact of backhaul limitations on the wide system performance. Focusing on the DL, the load on a backhaul link $j \in \mathcal{B}_h$ consists of the sum of downlink loads (corresponding to best-effort traffic) of all BSs using that link:

$$\sum_{i \in \mathcal{B}(j)} \rho_i^{D,b} \tilde{c}_i^D, \tag{7}$$

where $\tilde{c}_i^D$ is an estimate of the downlink total rate delivered by BS i. A BS is usually characterized by its "peak" rate (often upper bounded by the maximum MCS available), and a "busy" rate, when a BS serves many users [18]. 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 could measure this load and use it directly.

(C.3 - Backhaul provisioning) We have derived the backhaul link load ($\sum_{i \in \mathcal{B}(j)} \rho_i^{D,b} \tilde{c}_i^D$) and defined the backhaul capacity limitation ($C_h^D(j)$) for each backhaul link $j \in \mathcal{B}_h$ (see C.2). Thus, each of these links shall introduce a backhaul *constraint* to avoid exceeding its maximum capacity and prohibit backhaul congestion ($\sum_{i \in \mathcal{B}(j)} \rho_i^D \tilde{c}_i^D \prec C_h^D(j) \ \ \forall j \in \mathcal{B}_h$).

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 *provisioned* otherwise. We investigate the user-association problem separately for each scenario in Sections 3 and 4, by focusing on different tradeoffs.

3 User-Association for Provisioned Backhaul Networks

We start our discussion for optimal user-association by assuming that the backhaul network is provisioned and so, we can safely ignore it while deriving the optimal association rules. Our aim is to focus on the radio access network performance and traffic-differentiation involved tradeoffs.

We remind to the reader that based on our system model, the association policy consists in finding appropriate values for the routing probabilities $p_i^{l,t}(x), \ l \in \{D, U\}, \ t \in \{b, d\}$, for DL and UL, best-effort and dedicated traffic, respectively (defined earlier in assumption B.5 and B.9). That is, for each location x, we would like to optimally choose to which BS i to route different flow types generated from (UL) or destined at (DL) users in x[6]. Our goal for this association problem is

[6] The use of a probabilistic association rule simplifies solving the problem. As it will turn out, the optimal values will be either 0 or 1 (deterministic).

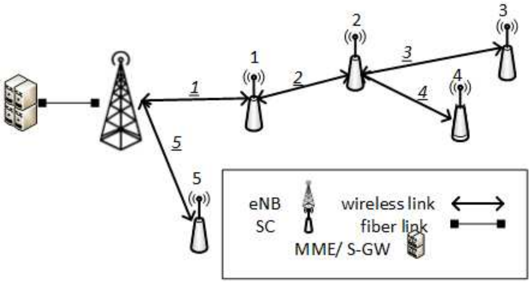


Figure 2: Backhaul topology in future Hetnet.

threefold: (i) ensure that the capacity of no BS is exceeded (later in Section 4, we will also include the constraint of no backhaul capacity is exceeded); (ii) achieve a good tradeoff between user physical data rates, user QoS and load balancing, (iii) investigate how *UL/DL association split* impacts the optimal rule derivation and the performance benefits of split UL/DL.

We define the feasible region for the aforementioned routing probabilities, by requiring that no BS capacity being exceeded.

Definition 1. *(Feasibility): Let $l \in \{U, D\}, t \in \{b, d\}$, and let ϵ be an arbitrarily small positive constant. The set $f^{l,t}$ of feasible BS loads $\rho^{\mathbf{l,t}} = (\rho_1^{l,t}, \rho_2^{l,t}, \ldots, \rho_{\|\mathcal{B}\|}^{l,t})$ is*

$$f^{l,t} = \Big\{ \rho^{\mathbf{l,t}} \mid \rho_i^{l,t} = \int_{\mathcal{L}} p_i^{l,t}(x) \rho_i^{l,t}(x) dx, \\ 0 \leq \rho_i^{l,t} \leq 1 - \epsilon, \\ \sum_{i \in \mathcal{B}} p_{i,t}^{l}(x) = 1, \\ 0 \leq p_{i,t}^{l}(x) \leq 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}. \tag{8}$$

Lemma 3.1. *The feasible sets $f^{D,b}, f^{D,d}, f^{U,b}, f^{U,d}$ as well as the $[f^{D,b}; f^{D,d}]$, $[f^{U,b}; f^{U,d}]$, $[f^{D,b}; f^{U,b}]$, $[f^{D,b}; f^{D,d}; f^{U,b}; f^{U,d}]$, are convex.*

Proof. The proof for the feasible set $f^{D,b}$ is presented in [9]. It can be easily adapted for the other cases, too (e.g., see [35]). □

3.1 Optimal Split UL/DL User Association

We first define the user association problem for the split UL/DL case. Here, we should require that all DL best-effort and dedicated flows at x have to be downloaded from the same BS, i.e., $p_i^D(x) = p_i^{D,b}(x) = p_i^{D,d}(x)$. Also, that all UL best-effort and dedicated should be offloaded to the same BS, so $p_i^U(x) = p_i^{U,b}(x) = p_i^{U,d}(x)$. Note that, $p_i^D(x)$ and $p_i^U(x)$ can take different values (see B.10) in split

UL/DL scenarios. Hence, the problem of optimal DL and UL association *can be decoupled into two independent problems, one for DL and one for UL*. In the remainder of this section, we focus on the optimal DL association problem, and we omit the superscripts $\{D, U\}$ to simplify notation. We return to the joint UL/DL association problem in the next subsection.

Following [9], we extend the α-cost function to consider performance for dedicated flows, along with the best-effort ones. Specifically, we introduce the parameter $0 \le \theta \le 1$ that linearly weights the relative importance between best-effort and dedicated traffic, and parameters $\alpha^b \ge 0$, $\alpha^d \ge 0$ that define the load balancing degree for the corresponding resources.

Theorem 3.2. *[Split UL/DL User Association rule] The optimal user-association problem can be expressed as* $\min_\rho\{\phi(\rho)|\rho = (\rho^{\mathbf{b}}; \rho^{\mathbf{d}}) \in f = (f^b; f^d)\}$*, where*

$$\phi(\rho) = \sum_{i\in\mathcal{B}} \theta \frac{(1-\rho_i^b)^{1-\alpha^b}}{\alpha^b - 1} + (1-\theta)\frac{(1-\rho_i^d)^{1-\alpha^d}}{\alpha^d - 1}, \textit{if } \alpha^d, \alpha^d \neq 1. \tag{9}$$

If the feasible domain f *of the problem is non-empty, and* $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ *denotes the optimal load vector, the user-association rule at location* x *is expressed by the following weighted harmonic mean (of individual rules) formula*

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b} \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b(x)\cdot\left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d(x)\cdot\left(1-\rho_i^{*b}\right)^{\alpha^b}} \tag{10}$$

where $e^b(x) = \frac{\theta z^D z^b}{\mu^b(x)c_i(x)}$ *and* $e^d(x) = \frac{(1-\theta)z^D z^d}{\mu^d(x)k_i(x)}$*, optimally weight the corresponding individual association rules depending on the traffic statistics.*

Note that if $\alpha^b = 1$ (or α^d=1) the corresponding term in the objective (Eq. 9) is not defined, and $\log(1-\rho_i^b)^{-1}$ ($\log(1-\rho_i^d)^{-1}$) is used instead.

Proof. We prove that the above user-association rule (Eq. 10) indeed minimizes the cost function of Eq. (9). This problem is a convex optimization because its feasible set f is convex (see Lemma 3.1). Also, the objective function $\phi(\rho)$ is convex, due to the summation and linear combinations of the convex function $\phi(\rho)$ that is proven to be convex in [9]. Let $\rho^* = [\rho^{*b}; \rho^{*d}]$ be the optimal solution of Problem (9). Hence, it is adequate to check the following condition for optimality

$$\langle \nabla\phi(\rho^*), \Delta\rho^*\rangle \ge 0 \tag{11}$$

for all $\rho \in f$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by(10), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\Big\{i = \arg\max_{i\in\mathcal{B}} \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b} \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b(x)\cdot\left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d(x)\cdot\left(1-\rho_i^{*b}\right)^{\alpha^b}}\Big\}. \tag{12}$$

Then the inner product in Eq. (11) can be written as:

$$\begin{aligned}
&\langle \nabla\phi\left(\rho^*\right), \Delta\rho^* \rangle = \sum_{z=\{b,d\}} \frac{\partial\phi}{\partial\rho_z}\left(\rho^*\right)\left(\rho_z - \rho_z^*\right) \\
&= \frac{\partial\phi}{\partial\rho^b}(\rho^*)(\rho^b - \rho^{*b}) + \frac{\partial\phi}{\partial\rho^d}(\rho^*)(\rho^d - \rho^{*d}) \\
&= \theta\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^b)^{\alpha^b}}(\rho_i^b - \rho_i^{*b}) + (1-\theta)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^d)^{\alpha^d}}(\rho_i^d - \rho_i^{d*}) \\
&= \sum_{i\in\mathcal{B}} \frac{\theta\int\rho_i^b(x)(p_i(x)-p_i^*(x))dx}{(1-\rho_i^b)^{\alpha^b}} + \frac{(1-\theta)\int\rho_i^d(x)(p_i(x)-p_i^*(x))dx}{(1-\rho_i^d)^{\alpha^d}} \\
&= \int_L \lambda(x)\sum_{i\in\mathcal{B}}(p_i(x)-p_i^*(x))\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d} + e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}dx
\end{aligned} \tag{13}$$

where $e^b(x) = \frac{\theta z^{DL} z^b}{\mu^b(x)c_i(x)}$ and $e^d(x) = \frac{(1-\theta)z^D z^d}{\mu^d(x)k_i(x)}$. Note that,

$$\begin{aligned}
&\sum_{i\in\mathcal{B}} p_i(x)\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d} + e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \geq \\
&\quad \sum_{i\in\mathcal{B}} p_i^*(x)\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d} + e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}
\end{aligned} \tag{14}$$

holds because $p^*(x)$ in (12) is an indicator for the minimizer of $\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d}+e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}$. Hence, (11) holds. □

While θ linearly weights the best effort versus dedicated flow performance (see Eq. 9), the impact of α^b, α^d is not obvious. We now discuss their impact on the system performance and refer to [9], [36] for the respective proofs.

- *Spectral Efficiency Optimization:* $\alpha^b = 0$ maximizes the average physical rate for best-effort flows (defined in B.3), whereas $\alpha^d = 0$ maximizes the average dedicated servers for dedicated flows (defined in B.8). Obviously, these optimize the user $SINR$ and spectral efficiency.

- *Optimizing related QoS metrics:* if $\alpha^b = 1$ the corresponding optimal rule tends to maximize the average user throughput. If $\alpha^b = 2$ the per-flow delay is minimized since the objective for best effort flows corresponds to the delay of an M/G/1/PS system. If $\alpha^d = 1$ the corresponding optimal rule becomes equivalent to the average *idle* dedicated servers in a k-loss system, and the actual blocking probability is minimized.

- *Load-Balancing Efficiency Optimization:* As $\alpha^b \to \infty$, we minimize the maximum BS utilization, i.e. load balancing between the ρ^b is achieved. Similar for α^d and ρ^d's. Note that, the point of α^b that all BS best-effort utilizations are equalized might be different from the one for dedicated, depending on the respective traffic statistics.

The above Theorem, defines in Eq.(10) the optimal association rule for each user at location x, given the optimal BS load vector ρ^*. However, as the optimal vector ρ^* is not necessarily known, in Section 6 we propose an *iterative* algorithm, that starts within a feasible load vector point, and through an iterative procedure it converges to the optimal one.

In the case of split UL/DL, the above analysis can be applied *separately* on UL and DL traffic, and optimize UL and DL associations independently. Or equivalently, optimize the arithmetic mean (or, sum) of the corresponding rules.

3.2 Optimal Joint UL/DL User Association

Current cellular networks (e.g. 3G/4G) suggest that a UE should be connected to a single BS for both UL and DL traffic [37]. This changes the optimal association problem, as one now needs to *jointly* optimize UL and DL traffic performance. E.g., a user at location x might end up being associated with a BS offering suboptimal performance on both the downlink and uplink, because other BS candidates offer really bad UL (or really bad DL) performance.

We thus need to modify our framework accordingly. First, while deriving the association rules we will have to require $p_i^D(x) = p_i^U(x)\ \forall i \in \mathcal{B}$. Second, UL and DL performance must now be included in the same cost function. Specifically, the operator may linearly weigh the importance of DL and UL traffic performance with a parameter $\tau \in [0,1]$[7].

Theorem 3.3. *[Joint UL/DL User Association rule] The optimal association problem can be expressed as* $\min_\rho \left\{ \phi(\rho) | \rho = [\rho^{\mathbf{D,b}}; \rho^{\mathbf{D,d}}; \rho^{\mathbf{U,b}}; \rho^{\mathbf{U,d}}] \in f = (f^{D,b}; f^{D,d}; f^{U,b}; f^{U,d}) \right\}$, *where*

$$\begin{aligned}\phi(\rho) = \tau \left(\sum_{i \in \mathcal{B}} \theta^D \frac{(1-\rho_i^{D,b})^{1-\alpha^{D,b}}}{\alpha^{D,b}-1} + (1-\theta^D) \frac{(1-\rho_i^{D,d})^{1-\alpha^{D,d}}}{\alpha^{D,d}-1} \right) + \\ (1-\tau) \left(\sum_{i \in \mathcal{B}} \theta^U \frac{(1-\rho_i^{U,b})^{1-\alpha^{U,b}}}{\alpha^{U,b}-1} + (1-\theta^U) \frac{(1-\rho_i^{U,d})^{1-\alpha^{U,d}}}{\alpha^{U,d}-1} \right).\end{aligned} \tag{15}$$

If the feasible domain f of the problem is non-empty, and given the set of all flow-types $\Omega = \{(D,b),(D,d),(U,b),(U,d)\}$, the optimal user-association rule at location x is now

$$i(x) = \arg\max_{i \in \mathcal{B}} \frac{\prod\limits_{c \in \Omega} ((1-\rho^{*c})^{\alpha^c})}{\sum\limits_{c \in \Omega} e^c(x) \prod\limits_{l \in \Omega \neq c} ((1-\rho^{*c})^{\alpha^c})}, \tag{16}$$

where $e^{D,b}(x) = \tau \frac{\theta^D z^D z^b}{\mu^{D,b}(x) c_i^D(x)}$, $e^{D,d}(x) = \tau \frac{(1-\theta^D) z^D z^d}{\mu^{D,d}(x) k_i^D(x)}$, $e^{U,b}(x) = (1-\tau) \frac{\theta^U z^U z^b}{\mu^{U,b}(x) c_i^U(x)}$ *and* $e^{U,d}(x) = (1-\tau) \frac{(1-\theta^U) z^U z^d}{\mu^{U,d}(x) k_i^U(x)}$ *are the corresponding weight factors.*

[7] If α^D or α^U is equal to 1, the respective fraction must again be replaced with $\log(1-\rho_i)$, as explained earlier.

Proof. We refer the interested reader to [36]. □

Remark 1. The above optimal rule derived in Eq. (16) suggests that in the *joint UL/DL* scenario associated with objectives that potentially conflict with each other (due to the different flow type performances), it is optimal to associate a user with the BS that maximizes a weighted version of the *harmonic mean* of the individual association rules when considering each objective alone. To better understand this, we focus on a simple scenario with only DL and UL best-effort traffic. And assume the following BS options for a user: (BS A) offers 50Mbps DL and only 1Mbps UL; (BS B) 200Mbps DL and 0.5Mbps UL; (BS C) 20Mbps DL and 5Mbps UL. If we care about UL and DL performance equally (i.e. $\tau = 0.5$), one might assume that the BS that maximizes the arithmetic mean (or arithmetic sum) of rates would be a fair choice (i.e. BS B). However, this would lead to rather poor UL performance. Maximizing the harmonic mean would lead to choosing (BS C) instead[8]. Additionally, note that in the case of *split UL/DL*, covered in Section 3.1, where each user is free to be associated with two different BSs for the DL and UL traffic offloading, DL traffic would be associated with (BS B), and UL traffic with (BS C) by maximizing the arithmetic mean (or, sum) of their throughputs [9]. These simple examples intuitively explain how split UL/DL impacts the user association policies, by allowing to independently optimize each objective. This also demonstrates why UL/DL split may perform considerably better than the joint association. We will further explore this in the simulations (Section 5).

We finally underline that, the “formula" of harmonic or arithmetic mean maximization further allows to add more dimensions in our setup and *flexibly* derive the optimal rules without any analytical calculations. For instance, consider a more modern offloading technique, where different downlink, or uplink, flow types are able to be offloaded to different BSs (e.g., per flow/QCI offloading) with conflicting aims. Using our model we can consider an additional respective α-function for each flow type, and either analytically or flexibly, optimize the complete objective as showed earlier.

4 User-Association for Under-Provisioned Backhaul Networks

While the rules derived above, that try to reflect different performance tradeoffs, always lead to BS loads that are supported from the access network, they perhaps will not be supported from the backhaul link (or the corresponding backhaul link path) for that BS, since they ignore potential backhaul limitations. To

[8] While this simple example captures the main principle, the actual rule is more complex, as it weighs each objective with the complex factor $e^l(x)$.

[9] The usage of harmonic mean and arithmetic mean/sum appears in a number of physical examples, such as in the calculation of the total resistance in circuits where all resistances are set in series or in parallel.

that end, in this section we try to extensively consider the backhaul network and related limitations while extracting the optimal association rules, and include to our goals (i) that no backhaul link is congested, (ii) the impact of backhaul topology and capacity on key performance metrics. In order to better elucidate this problem at hand and without loss of generality, we focus on a simple scenario with *only best-effort traffic*. So, in the remainder of the section we drop the corresponding superscripts “b", “d" to simplify notation.

One of the main challenges when attempting to consider these backhaul constraints is to maintain the user association policy distributed (famous solvers for such convex problems, e.g. through the Lagrangian dual function [38], require a centralized controller entity); in Section 6 we highlight why distributiveness is important. To that end, we chose to consider the backhaul constraints in the objective function as appropriate *penalty functions* [35]. This not only facilitates deriving a distributed implementation of the policy, but also allows us to treat the backhaul constraint as a “soft” constraint that ends up being “hard" and satisfy convergence to a feasible solution, as we shall see later.

4.1 Optimal Split UL/DL User Association

We follow the same presentation as the provisioned case, and start out discussion, with the split UL/DL case. As the association problem can be decoupled, in that case, into two independent problems, we focus on the optimal DL association problem, and we omit the superscripts $\{D, U\}$. We return to the joint UL/DL association problem in the next section. To better illustrate our approach, we first apply this for a simple star BH topology, and then generalize for a tree BH topology).

Optimal User Association for Star BH Topology)

In the following, since for star topologies there is exactly one backhaul link (j) associated with each BS (i), it is $i = j$ (see also C.1). Let $\mathcal{I}(i)$ be an indicator variable, that shows whether the i-th backhaul link is congested ($\mathcal{I}(i)$=1) or not ($\mathcal{I}(i)$=0). Precisely (see C.2)

$$\mathcal{I}(i) = \begin{cases} 0, & \text{when } \frac{\rho_i \tilde{c}_i}{C_h(i)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{17}$$

Theorem 4.1 (Split UL/DL User Association rule in a star BH topology)**.** *The optimal user-association problem with a star BH topology is expressed as* $\min_\rho \Big\{\phi(\rho) | \rho \in f\Big\}$*, where*

$$\phi(\rho) = \sum_{i \in \mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} + \gamma \sum_{i \in \mathcal{B}_h} \mathcal{I}(i) \left(\frac{\rho_i \tilde{c}_i}{C_h(i)} - 1 \right)^2. \tag{18}$$

If the feasible domain f of the problem is non-empty, and $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ denotes the optimal load vector, the user-association rule at location x is

$$\arg\max_{i\in\mathcal{B}} c_i(x) \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{C_h(i)} - 1\right)}. \tag{19}$$

Proof. We now prove that the above rule indeed minimizes the cost function of Eq. (18). This minimization is a convex optimization problem. Its feasible set f is convex, and the objective $\phi(\rho)$ is also convex due to the summation of two convex terms: the first is convex as discussed earlier, and the second due to the composition property of convexity [38]. Let ρ^* be the optimal solution of this minimization problem. Again, it is adequate to check for optimality if

$$\langle \nabla\phi(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{20}$$

for all $\rho \in f$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by (19), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{ i = \arg\max_{i\in\mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{C_h(i)} - 1\right)} \right\}. \tag{21}$$

Before proceeding to the calculation of the inner product, we analytically calculate the derivative of the corresponding cost function $\phi(\rho)$, described in Eq. (18). The derivative is an i-th dimensional vector; the i-th element of which has value:

$$\nabla\phi(\rho_i) = \begin{cases} (1-\rho_i)^{-\alpha}, & \text{if } \frac{\rho_i \tilde{c}_i}{C_h(i)} \leq 1 \\ (1-\rho_i)^{-\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i\tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2}, & \text{if } \frac{\rho_i \tilde{c}_i}{C_h(i)} \geq 1. \end{cases} \tag{22}$$

When $\rho_i = \frac{C_h(i)}{\tilde{c}_i}$, we work out explicitly from the definition to calculate the derivative. It is:

$$\lim_{\rho_i \to \frac{C_h(i)}{\tilde{c}_i}^+} \nabla\phi(\rho_i) = \lim_{\rho_i \to \frac{C_h(i)}{\tilde{c}_i}^-} \nabla\phi(\rho_i) = (1-\rho_i)^{-\alpha}. \tag{23}$$

Summarizing, the i-th element of the derivative of the considered function can be written:

$$\nabla\phi(\rho_i) = (1-\rho_i)^{-\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i\tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2}. \tag{24}$$

To that end, the inner product defined in Eq. (20), becomes:

$$\begin{aligned} \langle \nabla\phi\left(\rho^*\right), \Delta\rho^* \rangle &= \sum_{i\in\mathcal{B}} \left\{ \frac{1}{(1-\rho_i^*)^\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i^*\tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2} \right\} (\rho_i - \rho_i^*) \\ &= \sum_{i\in\mathcal{B}} \frac{1 + 2\gamma\mathcal{I}(i)(1-\rho_i^*)^\alpha \frac{(\rho_i^*\tilde{c}_i^2 - \tilde{c}_i C_h(i))}{C_h(i)^2}}{(1-\rho_i^*)^\alpha} \int_{\mathcal{L}} \rho_i(x)\left(p_i(x) - p_i^*(x)\right) dx \\ &= \int_L \frac{\lambda(x)}{\mu(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + \frac{2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \mathcal{I}(i)}{C_h(i)} \left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)} - 1\right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) \left(p_i(x) - p_i^*(x)\right) dx \end{aligned}$$

Note that,

$$\sum_{i\in\mathcal{B}} p_i(x)\left\{\frac{1+\frac{2\gamma(1-\rho_i^*)^{\alpha}\tilde{c}_i\mathcal{I}(i)}{C_h(i)}\left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)}-1\right)}{c_i(x)(1-\rho_i^*)^{\alpha}}\right\} \geq$$
$$\sum_{i\in\mathcal{B}} p_i^*(x)\left\{\frac{1+\frac{2\gamma(1-\rho_i^*)^{\alpha}\tilde{c}_i\mathcal{I}(i)}{C_h(i)}\left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)}-1\right)}{c_i(x)(1-\rho_i^*)^{\alpha}}\right\}$$

holds because $p_i^*(x)$ in (21) is an indicator for the minimizer of $\frac{1+2\gamma\cdot(1-\rho_i^*)^{\alpha}\cdot\tilde{c}_i\cdot\frac{\mathcal{I}(i)}{C_h(i)}\cdot\left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)}-1\right)}{c_i(x)(1-\rho_i^*)^{\alpha}}$. Hence (20) holds. □

We again expressed the objective (Eq. (18)) with respect to the variables ρ_i, for convenience. The first sum is the standard α-cost function for each BS i, already analyzed in the previous section. The second sum introduces a penalty for each backhaul link i whose capacity is exceeded ($\mathcal{I}(i) = 1$). This penalty function is quadratic on the amount of excess load (quadratic penalty functions are often considered in convex optimization literature [39]). We chose to solve the problem iteratively, by starting with a small constant for γ, according to the magnitude of the main cost function, that introduces a “soft" constraint (i.e., backhaul capacity can be “slightly” violated if this really improves the radio access performance). Then, using increasing γ values it eventually converges to a “hard” constraint (no violations are allowed), as usually done in optimizations based on penalty functions, in order to ensure that the algorithm doesn’t get stuck in steep valleys [39].

Regarding the optimal association rule of Eq. (19), we note that when the capacity constraint for the backhaul link i is not active (i.e., $\mathcal{I}(i) = 0$, in provisioned BH networks), the above theorem states that the optimal association rule is the same as the one found in [9], or the one defined in Eq. (10) when $\theta \to 1$. However, when the backhaul link of BS i gets congested, a second term is added in the denominator that penalizes that BS making it less preferable to UEs at location i, even if the offered radio access rate $c_i(x)$ is high, or the radio interface of i is not itself congested.

Optimal User Association for Tree BH Topology)

We now consider a more complex backhaul scenario, where a single backhaul link might route traffic from multiple BSs, and the traffic of a single BS might be routed over multiple backhaul links (multi-hop path) towards the eNB. $\mathcal{I}(j)$ is now

$$\mathcal{I}(i) = \begin{cases} 0, & \text{when } \frac{\sum_{i\in\mathcal{B}(j)}\rho_i\tilde{c}_i}{C_h(j)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{25}$$

Theorem 4.2. *[Split UL/DL User Association rule in a tree BH topology] The optimal user association problem with a tree BH topology is expressed as* $\min_{\rho}\Big\{\phi(\rho)|\rho \in$

$f \Big\}$, *where*

$$\phi(\rho) = \sum_{i \in \mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} + \gamma \sum_{j \in \mathcal{B}_h} \mathcal{I}(j) \left(\frac{\sum\limits_{i \in \mathcal{B}(j)} \rho_i \tilde{c}_i}{C_h(j)} - 1 \right)^2 . \tag{26}$$

If the feasible domain f of the problem is non-empty, the optimal user-association rule at location x is now

$$\arg\max_{i \in \mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{C_h(j)} \cdot \left(\frac{\sum_{k \in \mathcal{B}(j)} \rho_k^* \tilde{c}_k}{C_h(j)} - 1 \right)} . \tag{27}$$

Proof. The steps of this proof are similar to the star case, so we present here directly the corresponding inner product.

$$\begin{aligned}
&\langle \nabla \phi(\rho^*), \Delta \rho^* \rangle = \\
&= \sum_{i \in \mathcal{B}} \Big\{ \frac{1}{(1-\rho_i^*)^\alpha} + 2\gamma \sum_{j \in \mathcal{B}_h(i)} \mathcal{I}(j) \Big[\frac{\sum_{k \in \mathcal{B}(j)} \rho_k^* \tilde{c}_k}{C_h(j)^2} \tilde{c}_i - \frac{\tilde{c}_i}{C_h(j)} \Big] \Big\} (\rho_i - \rho_i^*) \\
&\quad \cdot \int_{\mathcal{L}} \rho_i(x) \, (p_i(x) - p_i^*(x)) \, dx = \\
&= \int_L \frac{\lambda(x)}{\mu(x)} \sum_{i \in \mathcal{B}} \left(\frac{1 + 2\gamma (1-\rho_i^*)^\alpha \tilde{c}_i \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{C_h(j)} \cdot \left(\frac{\sum\limits_{k \in \mathcal{B}(j)} \rho_k^* \tilde{c}_k}{C_h(j)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) . \\
&\quad \cdot (p_i(x) - p_i^*(x)) \, dx \geq 0,
\end{aligned} \tag{28}$$

due to the corresponding minimizer $p_i^*(x)$ derived from (27). □

As one can see, the cost function is similar in nature. The first term corresponding to the radio access part remains unchanged. The second term again introduces a penalty for each backhaul link that is congested. However, there are a number of interesting differences between the star and tree cases. First, the penalty term in the denominator of the optimal association rule (Eq. (27)) now considers the whole backhaul path $\mathcal{B}_h(i)$ that traffic from BS i traverses, and adds a penalty for *every* link along that path that is congested (outer sum in the denominator). This observation provides some support for the number of backhaul hops heuristic proposed in [27, 40]. However, our analysis also suggests that it can be suboptimal, as a path with few hops might still include one or more congested links, and provides the optimal way to weigh in the amount of congestion on each backhaul link.

Second, the actual congestion on each backhaul link j is now not only dependent on the load of the candidate BS i, but also on other BSs whose load is routed over j. Hence, a BS i which would otherwise be a good candidate for traffic at location x, might still be penalized and not selected, even if it does not impose itself a large load on a backhaul link j. This is because *other* BSs sharing the same backhaul link might be heavily loaded or congested.

In the case of split UL/DL traffic, the above analysis can be applied *separately* on UL and DL traffic, and optimize UL and DL associations independently. Finally, although we have provided separate solutions for star and tree topologies, to better illustrate our approach, the optimal rule for the tree topology is generic, and includes star topologies as well.

4.2 Optimal Joint UL/DL User Association

Here, we need to modify our framework accordingly, as we did in Section 3.2, to include (i) that $p_i^D(x) = p_i^U(x)\ \forall i \in \mathcal{B}$, (ii) the weigh of importance between DL and UL traffic performance $\tau \in [0, 1]$[10]. If $\rho = [\rho^D; \rho^U] \in f = \{f^D; f^U\}$, our objective now is

$$\phi(\rho) = \sum_{i\in\mathcal{B}} \tau \frac{(1-\rho_i^D)^{1-\alpha^D}}{\alpha^D - 1} + (1-\tau)\frac{(1-\rho_i^U)^{1-\alpha^U}}{\alpha^U - 1}, \text{if } \alpha^D, \alpha^U \neq 1. \tag{29}$$

We also need to extend the penalty function to consider both uplink and downlink capacity being exceeded on the backhaul link. Here, we present our results directly for the general case of tree backhaul topology, and we remind the reader that this is applicable to star backhaul topologies as well.

Theorem 4.3 (Joint UL/DL User Association rule in a tree BH topology)**.** *The optimal association problem with a generic BH topology is expressed as* $\min_\rho \Big\{\phi(\rho) | \rho = [\rho^{\mathbf{D}}; \rho^{\mathbf{U}}] \in f\Big\}$*, where*

$$\phi(\rho) = \phi(\rho) + \gamma \sum_{k\in\{D,U\}} \sum_{j\in\mathcal{B}_h} \mathcal{I}^k(j) \left(\frac{\sum\limits_{i\in\mathcal{B}(j)} \rho_i^k \tilde{c}_i^k}{C_h^k(j)} - 1 \right)^2 . \tag{30}$$

If the feasible domain f of the problem is non-empty, the optimal user-association rule at location x is

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{\left(1-\rho_i^{*D}\right)^{\alpha^D} \cdot \left(1-\rho_i^{*U}\right)^{\alpha^U}}{e^D(x)\cdot\left(1-\rho_i^{*U}\right)^{\alpha^U} + e^U(x)\cdot\left(1-\rho_i^{*D}\right)^{\alpha^D}}, \tag{31}$$

where if $g^D = \tau, g^U = 1-\tau$, then for $l \in \{D, U\}$:

$$e^l(x) = \frac{z^l \left(g^l + 2\gamma \left(1-\rho_i^{*l}\right)^{\alpha^l} \sum\limits_{j\in\mathcal{B}_h(i)} \frac{\mathcal{I}^l(j)}{C_h^l(j)} \left(\frac{\sum\limits_{k\in\mathcal{B}(j)} \rho_k^{*l} \tilde{c}_k^l}{C_h^l(j)} - 1 \right) \right)}{\mu^l(x) c_i^l(x)}.$$

[10] If α^D or α^U is equal to 1, the respective fraction must again be replaced with $\log(1-\rho_i)$, as explained earlier.

Proof. We refer the interested reader to [35]. □

The penalty function for the backhaul network is simply the sum of the respective penalty functions for UL and DL, described in Theorem 4.1. However, despite the similarities of the cost functions, as we can see, the resulting association policy in the joint UL/DL case is more complex.

For completeness, we present the optimal user-association rule in case of dedicated traffic too, for the joint UL/DL association. Here, it is adequate to “split" the DL backhaul resources of the j-th link ($C_h^D(j)$) between DL best-effort ($C_h^{D,b}(j)$) and DL dedicated traffic ($C_h^{D,d}(j)$), and treat each load of these “pipes" as a certain backhaul constraint with further appropriate penalty functions, as showed previously. Similarly in the UL scenario.

Theorem 4.4. *Given the set of all flow-types $\Omega = \{(D,b),(D,d),(U,b),(U,d)\}$ we present the optimal association rule at location x*

$$i(x) = \arg\max_{i \in \mathcal{B}} \frac{\prod\limits_{c \in \Omega} \left((1-\rho^{*c})^{\alpha^c}\right)}{\sum\limits_{c \in \Omega} e^c(x) \prod\limits_{l \in \Omega \neq c}^{4} \left((1-\rho^{*c})^{\alpha^c}\right)}, \tag{32}$$

where the complex factors $e^{l,b}(x)$, for $l \in \{D,U\}$ are

$$\frac{z^l z^b \left(g^l \theta^l + 2\gamma \left(1-\rho_i^{*l,t}\right)^{\alpha^{l,b}} \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}^{l,b}(j)}{C_h^{l,b}(j)} \left(\frac{\sum_{k \in \mathcal{B}(j)} \rho_k^{*l,b} \tilde{c}_k^{l,b}}{C_h^{l,b}(j)} - 1 \right) \right)}{\mu^{l,b}(x) c_i^l(x)},$$

whereas the $e^{l,d}(x)$, for $l \in \{D,U\}$ are

$$\frac{z^l z^d \left(g^l (1-\theta)^l + 2\gamma \left(1-\rho_i^{*l,d}\right)^{\alpha^{l,d}} \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}^{l,d}(j)}{C_h^{l,d}(j)} \left(\frac{\sum_{k \in \mathcal{B}(j)} \rho_k^{*l,d} \tilde{c}_k^{l,d}}{C_h^{l,d}(j)} - 1 \right) \right)}{\mu^{l,d}(x) k_i^l(x)}.$$

5 Simulations

In this section we briefly present some numerical results and discuss related insights. We consider a 2×2 km^2 area. Figure 3(a) shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ ($flows/hour$ per unit area) (blue implying low traffic and red high), with 2 hotspots. We assume that this area is covered by two macro BSs and eight SCs. The macro BSs that are shown with asterisks are numbered from 1-2, and the SCs that are shown with triangles are numbered from 3-10, as we can see in Fig. 3(b)-(c), Fig. 4, and in Fig. 5. We also consider standard parameters as adopted in 3GPP [41], listed in Table 1[11]. If not explicitly

[11] As for (i) the sizes and ratios of different flows, (ii) splitting parameters, we can use different values in order to capture different simulation scenarios, and derive similar results.

Table 1: Simulation Parameters

Parameter	**Variable**	**Value**
Transm. Power of eNB/ SC/ UE	$P_{eNB}/P_{SC}/P_{UE}$	43/24/12 dBm
BS Bandwidth for DL, UL	w/W	10/10 MHz
Noise Power Density	N_0	-174 dBm/Hz
Splitting parameter for DL, UL	ζ_i^D, ζ_i^U	0.5/0.5
Average DL/UL flow sizes	$\frac{1}{\mu^{D,b}}/\frac{1}{\mu^{U,b}}$	100/20 Kbytes
Average DL/UL flow demands	$B^D(x)/B^U(x)$	512, 128 kbps
Different flow ratios	z^b, z^D	0.3,0.6

mentioned, we assume $\theta^D = \theta^U = \tau = 0.5$, and the split UL/DL scenario as default.

Before proceeding, we need to setup a metric to evaluate load balancing (or, utilization) efficiency. Thus, we introduce the Mean Squared Error ($MSE^{D,b}$), between the DL best-effort utilization of different BSs, normalized to 1:

$$\text{MSE}^{D,b} = \frac{1}{2 \cdot \left\lfloor \frac{\|\mathcal{B}\|}{2} \right\rfloor \cdot \left\lceil \frac{\|\mathcal{B}\|}{2} \right\rceil} \sum_i \sum_j (\rho_i^{D,b} - \rho_j^{D,b})^2. \tag{33}$$

We define the DL load balancing metric for best-effort traffic to be $1 - MSE^{D,b}$, that increases on the amount of load balancing[12]. Similarly, we can define them for the other three cases $1 - MSE^{D,d}$, $1 - MSE^{U,b}$, $1 - MSE^{U,d}$.

5.1 Provisioned Backhaul

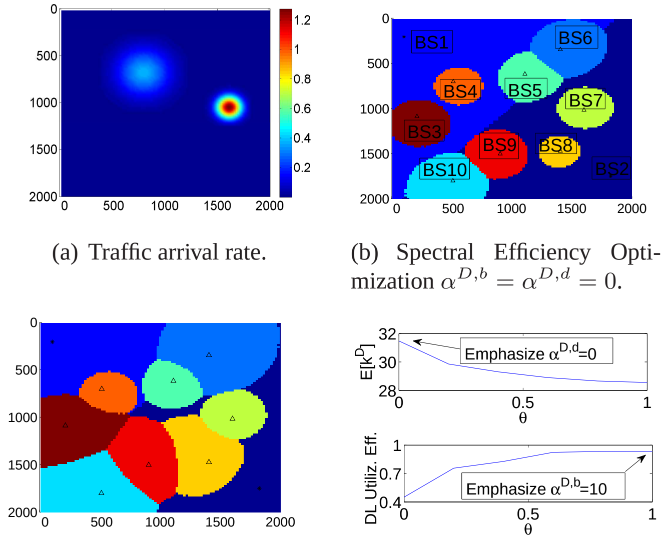
We now focus on the case of provisioned backhaul as considered in Section 3 and investigate the involved tradeoffs both qualitatively and quantitatively. We will present the impact of our proposed association rules via coverage snapshots to show how users associate in the considered network, while we will also provide values for related performance metrics that complete our study numerically.

Spectral efficiency vs. Load balancing. Figure 3(b) outlines the optimal DL user-associations if $\alpha^{D,b} = \alpha^{D,d} = 0$, i.e., when *spectral efficiency* is maximized. Thus, each UE at x is attached to the BS that offers the *highest DL SINR* and promises higher DL physical rate for best effort flows $c_i^{D,b}(x)$, and more “dedicated" servers $k_i^D(x)$; i.e. most of UEs are attached to macro BSs due to their high power transmission, and fewer to SCs, forming small circles around them. Consequently, macrocells are overloaded and load imbalance within the cells is sharpened (decreased $1-MSE^{D,b}$, $1-MSE^{D,d}$; see line 1 of Table 2). However, in Fig. 3(c) we emphasize the *load-balancing* efficiency and set $\alpha^{D,b} = \alpha^{D,b} = 10$. Now, most

[12] We should note that different load balancing metrics could have been used, e.g. the *maximum, median and minimum* BS load; however, we chose to use MSE since it facilitates the visualization of the network efficiency.

SCs vastly increase their coverage area in order to offload the overloaded macro BSs (e.g., BSs 6, 8, 10); "heavily" loaded (due to the hotspots) BSs, roughly maintain the same coverage (BS 4 and 7). Thus load balancing is improved, at the cost of $E[c^{D,b}]$, $E[k^{D}]$ (see line 2 of Table 2). For further implications of α parameters we refer the reader to [9].

Best-effort versus dedicated traffic performance. Although in the previous scenarios the best-effort- and dedicated- related traffic rules (represented from $\alpha^{D,b}$, $\alpha^{D,d}$) are aligned, one could ask how would two conflicting optimization objectives affect our network? The answer lays in the usage of θ^{D}, that judges which objective carries more importance. E.g., an operator has two main goals: (i) to maximize the average number of servers for "dedicated" traffic captured by $E[k^{D}]$ (set $\alpha^{D,d} = 0$), (ii) to better balance the utilization of best-effort resources between BSs (set $\alpha^{D,b} = 10$). As shown in Fig. 3(d), if $\theta \to 0$ $E[k^{D}]$ is maximized, whereas as $\theta \to 1$, $1 - MSE^{D,b}$ (DL best-effort load balancing) is optimized, and each objective comes at the price of the other.



(a) Traffic arrival rate.

(b) Spectral Efficiency Optimization $\alpha^{D,b} = \alpha^{D,d} = 0$.

(c) Enhanced Load Balancing $\alpha^{D,b} = \alpha^{D,d} = 10$.

(d) Ded. vs. Best-effort flows performance.

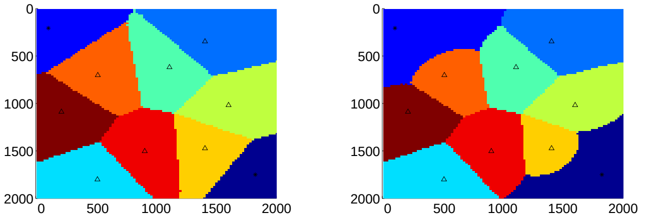
Figure 3: DL Optimal user-associations (Spectral efficiency vs. Load balancing and best-effort vs. ded. traffic performance)

DL vs. UL traffic performance is considered in Figure 3(b), 4(a)-4(b), with respective numerical performance metrics in Table 3. The first two figures depict the DL and UL optimal associations, in case of split UL/DL, for each user at x. However, if split is not available from the operator point of view, we have to weight whether the DL or UL performance is more important while selecting a *single* BS for joint UL/DL association, using parameter τ. To that end, Figure 3(b) (also)

Table 2: Numerical values for Figure 3.

	Rates and Servers		Load Balancing	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	1-$MSE^{D,b}$	1-$MSE^{D,d}$
Fig. 3(b)	16.3	32	0.77	0.78
Fig. 3(c)	14.3	27	0.96	0.995

outlines the optimal associations in the joint UL/DL case if the whole emphasis is on the *DL performance* ($\tau = 1$): this hurts the UL performance due to the asymmetric transmission powers of the UEs and BSs (see line 1 of Table 3). In Fig. 4(a) the emphasis is moved on the *UL performance* ($\tau = 0$), and each UE is attached to the nearest BS, in order to minimize the path loss [33] and enhance the UL performance; this hurts its DL performance though (see line 3 of Table 3). Finally, Fig. 4(b) shows the optimal coverage areas when one assigns equal importance to the UL and DL performance (i.e. $\tau = 0.5$): this moderates both DL and UL performance (line 2 of Table 3). This also corroborates the notion that split is able to simultaneously optimize UL and DL performances, as already discussed in theory.



(a) UL performance optimization $\tau = 0$.

(b) Equal emphasis of DL vs. UL $\tau = 0.5$.

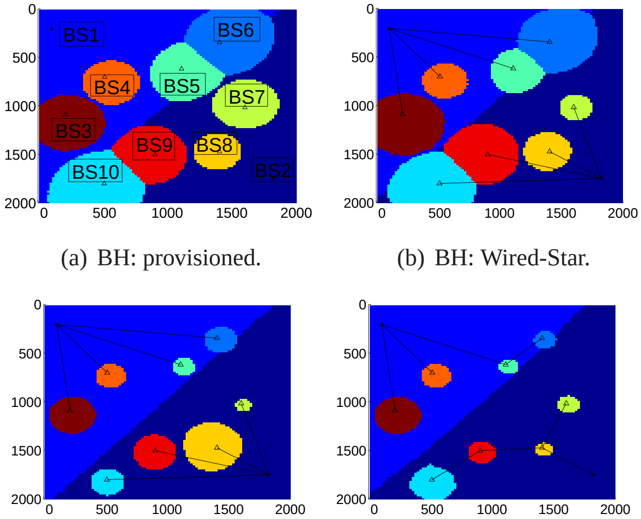
Figure 4: Optimal user-associations (DL vs. UL traffic performance)

Table 3: Numerical values for Figure 4.

	DL performance		UL performance	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	$E[c^{U,b}]$ (Mbps)	$E[k^U]$
Fig. 3(b)	16.3	32	2.3	18
Fig. 4(b)	14.7	28	3	24
Fig. 4(a)	13.3	26	3.6	28

5.2 Under-provisioned Backhaul

We now continue with some backhaul-limited network scenarios. We remind to the reader that our focus is on the backhaul links *between the macro cells and SCs* (for simplicity we assume provisioned links between the macro cells and core network). As already discussed in assumption C.1, we investigate two different backhaul topology families: (i) “star" topologies (single-hop paths), (ii) “tree" topologies (with multi-hop paths), along with two backhaul links types: *wired and wireless*[13]. Our aim is to evaluate the derived association rules described in Section 4 for different *under-provisioned* scenarios, by fixing the aforementioned trade-offs related to the traffic differentiation as it follows: $\theta^D = \theta^U = 1$ (we only focus on the best-effort flows by dropping the superscripts “b” and “d”), and $\alpha^D = \alpha^U = 1$ (throughput optimal values). Also, we assume *fixed* backhaul routing paths, pre-established with traditional Layer 2 routing, that the BH capacities on the DL and UL are the same (i.e. $C_h^D(j) = C_h^U(j) = C_h$, $\forall j \in \mathcal{B}_h$), and if not explicitly mentioned we assume them to be equal to $400Mbps$. We maintain this assumption to facilitate our discussion, although our framework works for heterogeneous backhaul links and UL/DL capacities (see C.2).



(a) BH: provisioned.

(b) BH: Wired-Star.

(c) BH: Wireless-Star.

(d) BH: Wireless-Tree.

Figure 5: DL optimal associations in different scenarios.

Before proceeding, we need to make an assumption about the backhaul link capacities. In case of *wired* backhaul links, we assume that the peak backhaul capacity C_h is always guaranteed. For *wireless* backhaul links we adopt a simple model associating peak backhaul capacity to distance: if the length of the i-th link is r_i, the peak capacity drops as:

$$d(r_i) = \begin{cases} 1, & r_i \leq r_0 \\ (\frac{r_0}{r_i})^n, & \text{otherwise}, \end{cases} \tag{34}$$

[13] Note that copper and fiber access are the key technologies for wired backhaul links, and microWave and millimeter-wave P2P or P2MP access are the counterpart for the wireless backhaul links [42].

where r_0 is some threshold range within which the maximal rate is obtained (e.g. Line-of-Sight), and n is the attenuation factor. Hence, the available capacity drops to $d(r_i)C_h(j)$ $(\leq\ C_h(j))$. For our simulations, we assumed that $r_0 = 200m$, and $n = 3$. While the above model is perhaps oversimplifying, our main goal is to simply include a generic model for the propagation related impact on wireless backhaul, compared to wired, without getting into the details of specific backhaul implementations. For detailed path loss models for different backhaul technologies, we refer the interested reader to [28].

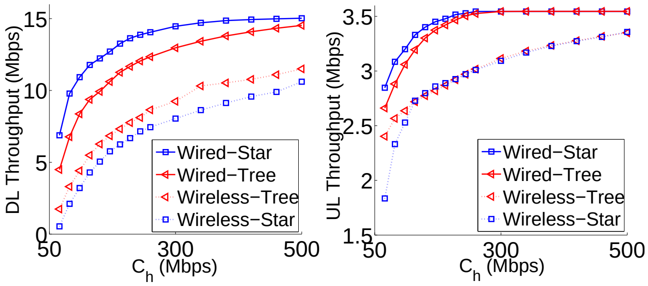
Coverage Snapshots. In Fig. 5(a) we depict the optimal DL user-associations for provisioned backhaul network with respect to the traffic arrival rates shown in Fig. 3(a). Compared to the associations showed in Figure 3(b) where $\alpha^D = \alpha^U = 0$, we note that now some SCs have slightly increased coverage area, in order to improve the mean user throughput [9].

In the following, we focus on different *under-provisioned* backhaul scenarios, and study the DL associations (similar behavior in the UL as explained in [35]). In Fig. 5(b) we adopt a *wired-star* backhaul topology, where SCs shrink their coverage areas, by handing-over users to other BSs, in order to offload the corresponding (under-provisioned) backhaul links; this phenomenon becomes more intense in the “hot-spot" areas (e.g., BS7 have vastly decreased their coverage areas) due to the higher traffic demand. Similarly, in Fig. 5(c), we assume a *wireless-star* backhaul topology, where SCs further decrease their coverage areas, due to the higher backhaul capacity loss caused from the long wireless links (see Eq.(34)).

In Fig. 5(d) we adopt a *wireless-tree* topology, where some SCs are required to carry also traffic of other SCs, and end up more congested. As a result, most SCs further decrease their coverage area, compared to the star-wireless topology. However, BS7 and BS10 enlarge their coverage areas, compared to the star case. This occurs because these SCs are far from the eNB, and multi-hop topology allows them to route their traffic over shorter wireless links with smaller capacity losses, compared to the star case (Fig. 5(c)). Hence, there are two main factors affecting the coverage areas in such wireless backhaul networks: (*topology*) each BS-load might traverse through multi-hop backhaul paths, by “wasting" resources from more than one backaul links (drawback for tree topologies); *(location)* the higher the η,r_0 the worse the capacity loss “wastage" over a dedicated direct backhaul link (drawback for star topologies that require longer links).

As backhaul networks become increasingly complex, e.g. “mesh" topologies, each BS has *multiple* possible routing paths to follow, beyond what is shown in the figures (we remind the reader that the above shown topologies are simply the given spanning routing trees). The above observations thus underline the shortcomings of predetermined, Layer 2 (L2) backhaul routing mechanisms, and call for a *joint* optimization of user-association on the radio access network along with dynamic, Layer 3 (L3) backhaul routing (see Section 6).

Under-provisioning impact on user performance. Figure 6(a), 6(b) depict the *average* DL and UL user throughputs, as a function of the backhaul capacity constraint C_h, on different scenarios. Generally, as C_h drops, the mean throughputs



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

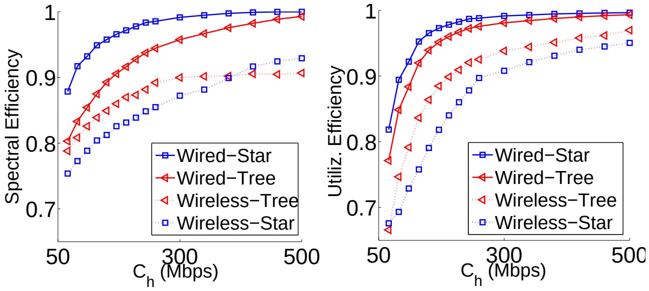
Figure 6: Mean throughputs overall all users in the network.

are decreased, since users are handed over to (potentially far-away) macro BSs, causing performance degradation. Interestingly, *the slope of the dropping rate* becomes more steep for lower values of C_h, due to the logarithimic capacity formula chosen in assumption (B.2). Also, as C_h increases, the average throughputs “converge" to the value corresponding to a provisioned backhaul network. Note that the average UL throughput convergences more quickly, compared to the DL. This happens due to the asymmetry between the DL and UL traffic demand on the radio access network: the UL one is much lower, mainly due to the asymmetry between the transmission powers of BSs and UEs, as well as different file sizes assumed in each direction. Beyond this point, the UL backhaul resources will be underutilized. This calls for a *flexible* TDD duplexing scheme, that will dynamically distribute the backhaul resources accordingly, for example by giving more backhaul resources to DL when the UL demand is already satisfied (e.g. the eIMTA scheme [43]). Finally, in the wired case, star topology is always slightly better than the tree, whereas in the wireless the opposite, as explained earlier.

Table 4: Mean throughp. for handed-over users (in Mbps).

Topology	$C_h = 50$	250	500 (Mbps)
DL / UL thr.: Star-Wired	1.1 / 0.2	3.1 / 1.6	4.1 / X
DL / UL thr.: Tree-Wired	0.6 / 0.1	2.4 / 0.7	3.2 / X
DL / UL thr.: Tree-Wirel.	0.2 / 0.03	1.7 / 0.07	2.1 / 0.15
DL / UL thr.: Star-Wirel.	0.1 / 0.001	1.4 / 0.05	1.7 / 0.02

One could notice that user throughputs drop slightly on the C_h constraint, e.g. in a wired-star topology if C_h drops $500 \rightarrow 50$ Mbps (10 times), the mean user throughput only drops $15 \rightarrow 6$ Mbps (~ 3 times). This is due to the fact that, under-provisioned backhaul links do not affect the whole network, but specific groups of users associated with the cells that suffer from low backhaul capacity. To better illustrate this, in Table 4 we show the average throughput of the *handed-over users*, as a function of C_h. Indeed, their performance is severely affected: for the same scenario, their DL throughput drops all the way to 1.1 Mbps ($\sim$ 15 times). (In scenarios with no handovers, we mark the respective table entry with an X.)



(a) DL Spectral efficiency.

(b) DL Load balancing (Utilization) Efficiency.

Figure 7: Downlink Network Efficiencies (normalized).

Under-provisioning impact on Network Performance. Turning our attention to network-related performance, Fig. 7(a) considers spectral efficiency ($bit/s/Hz$), *normalized* by the *maximum* corresponding value when the network is provisioned. Load-balancing (“utilization") efficiency is further considered in Fig. 7(b) in terms of the MSE metric, described earlier. Both efficiencies converge to 1 as the network gets provisioned. Low C_h values will push users to handover to far-away BSs, and this will potentially decrease their $SINR$ (spectral efficiency decrease), and create steep differences between BSs loads, e.g. by congesting macro BSs and under-utilizing the SCs (load balancing decrease). Note that, the joint degradation of these performances also impacts user performance negatively (e.g. user throughput), as explained in Section B.6. Regarding spectral efficiency, more specifically, although in the wired scenario, star topology is always better compared to the tree, in the wireless scenario this is not the case. For low values of C_h, the star topology is worse, due to the higher capacity loss of the long and direct links. However, as C_h is increased, and some links start becoming provisioned in the star topology, the capacity loss cost due to the long wireless links in the star topology, is dominated from the capacity loss cost due to multi-hop sharing links of the tree topology, by making tree a worse choice. We highlight that this trade-off can suggest different topologies as optimal in different under-provisioned scenarios, and can affect different performance metrics.

Table 5: UL/DL Split Vs. Joint-association Improvements

Performance	$\tau = 0$	$\tau = 0.5$	$\tau = 1$
DL / UL Throughput	6% / 32%	4% / 35%	0% / 37%
DL / UL Spectr. Eff.	4% / 29%	3% / 31%	0% / 33%
DL / UL Uiliz. Eff.	7% / 34%	4% / 38%	0% / 41%

Split UL/DL impact. As discussed earlier, while split is able to optimize the DL and UL performance, *simultaneously*, joint UL/DL association is incapable of this parallel optimization and using $0 \leq \tau \leq 1$ we can trade-off which dimension carries more importance. Table 5 illustrates the *performance improvements* that split promises over the joint UL/DL association, in terms of various metrics, for various τ when backhaul is underprovisioned. We underline that split enhances the UL performance considerably, e.g. the average UL throughput is increased up to 37%. This is due to the *dependency* that joint UL/DL generates between the DL and UL associations in the access network, that often makes the DL the bottleneck in the backhaul (due to aforementioned asymmetry between the peak access rates). Thus, DL will often “preempt" the backhaul constraint, and potentially (i) leave some UL resources unused, (ii) cause UL performance degradation.

6 Discussion and Future work

In this section, we complete our framework by proposing a distributed implementation. We also briefly discuss potential extensions of our framework, besides the “per-flow" offloading discussed at the end of Section 3.

Framework Implementation. There have been many efforts in the literature toward developing various *centralized* user association rules, to improve load balancing [44, 45]. These require a centralized controller entity that governs the BSs and the UEs with access to all the necessary information. However, depending on the operator capabilities such an implementation may not be applicable. Additionally, even when it is applicable, it may (a) require excessive message overhead and computational complexity that increase exponentially in the network size, as well as (b) allow only for slow adaptation on the queuing statistics at relatively long timescales, since such a controller is usually implemented in a server deep in the core network. Thus, to avoid relying on a centralized controller, current systems aim on distributed implementations.

Following [9], we sketch a distributed implementation that is applied iteratively, adapts to spatial traffic loads, and mainly involves two parts: the *user* and *base station* tier. At the k-th period, each user at some location x receives from different BSs the required value that relates to its both access and backhaul network performance in order to apply the association rule (e.g., in Eq. (19) this value corresponds to the fraction seen), e.g. through broadcast control messages[14]. Then each new flow request simply selects the BS i that maximizes the corresponding quantity. Also, at each k iteration, BSs measure their average utilizations $\rho^{(k)}$ after some required period of time (e.g., see Eq. (5)). Then, based on the previous BS loads $\tilde{\rho}^{(k)}$, the new BS load vector $\tilde{\rho}^{(k+1)}$ needed for the broadcast control message in the next iteration would be

$$\tilde{\rho}^{(k+1)} = \beta^{(k)} \cdot \rho^{(k)} + (1 - \beta^{(k)}) \cdot \tilde{\rho}^{(k)}, \tag{35}$$

[14] IEEE 802.16m facilitates these types of message structure [46].

where $\beta^{(k)} \in [0,1)$ is an exponential-averaging parameter. Note that, in the split UL/DL scenario, the UL and DL loads can be independently updated, whereas in the joint UL/DL should be updated jointly using the same $\beta^{(k)}$.

This iteration converges to the globally optimal point ρ^*, requiring a simple modification to the proof found [9]. Note that our framework could also be implemented in an SDN framework, using a centralized or hierarchical implementation, where a controller derives the optimal associations and directly sends them through the network to the UEs. We refer the interested reader to [36] [35] for such an implementation.

Dynamic TDD schemes on the access and backhaul networks. As discussed in assumption B.2, the (access) resource allocation between best-effort and dedicated traffic is applied according to a parameter ζ, whose optimization is out of the scope of this paper. Interestingly, one can include this resource allocation parameter ζ in the considered cost function, and attempt to tackle the complete problem by optimizing both parameters ρ and ζ, simultaneously. Specifically, the new cost function will now look like

$$\phi(\rho,\zeta) = \sum_{i\in\mathcal{B}} \theta \frac{(1-\frac{\rho_i^b}{\zeta_i})^{1-\alpha^b}}{\alpha^b - 1} + (1-\theta)\frac{(1-\frac{\rho_i^d}{1-\zeta_i})^{1-\alpha^d}}{\alpha^d - 1}, \text{if } \alpha^d, \alpha^d \neq 1. \tag{36}$$

Note that, the above objective is block separable, since for fixed ζ, it decomposes into two problems with optimization parameters ρ^b and ρ^d. Thus, it makes sense to decompose the objective into optimization levels, by following some well-known principles of *decomposition optimization* [47]. This provably reduces the algorithmic complexity and maintains our approach amenable to distributed implementations. Thus, at the lower level we have two subproblems, where in a fine timescale we attempt to derive the optimal value for the local variable $\rho = [\rho^d;\ \rho^b]$ for a fixed ζ, using the iterative methods described in this paper. In the higher level we encounter the master problem where we attempt to update the complicating variable ζ in a larger timescale (e.g., through the Newton method [38]), such that the overall objective described in Eq. (36) is improved, and we re-solve the two subproblems. This procedure is iterated until both local and complicating variables converge to their optimal values.

In simulations we showed that fixed split between UL and DL backhaul resources hurt performance; thus, a similar approach can also be taken to optimally allocate backhaul resources (see C.2). Finally, a *hierarchical decomposition algorithm* [47] could be used to *jointly* solve both the backhaul and radio access resource allocation (at slower time scales), together with the optimal user association problem.

Joint radio and L3 backhaul routing. Mesh backhaul topologies with multiple available routing paths are expected to be the rule, rather than the exception in future networks. Our assumption of fixed, L2 backhaul routing is restrictive, and as we saw in the simulations also penalizes performance. It would be interesting to jointly optimize (a) the BS that each user should be associated with, as well as (b) the routing path up to an aggregation point (L3 routing). Our goal is twofold: to consider (a) *per-BS offloading*, where each BS should offload all flows by using the same routing path upto an aggregation point, (b) *per-location offloading*, where flows at different locations of a certain BS can follow different routing paths to improve system performance. It remains to be investigated whether these two options retain the convexity and other desirable properties of the original problem.

7 Conclusion

In this paper, we propose a user-association framework for future HetNets by investigating both (a) provisioned, and (b) underprovisioned backhaul network scenarios. We showed how traffic differentiation, different backhaul topologies and capacity limitations affect the user and network performance, with joint consideration of the access and backhaul resources. Initial simulation results corroborate the correctness of our framework, and reveal interesting tradeoffs for different network scenarios, as well as potential drawbacks of schemes operated in the backhaul, currently.

References

[1] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, “Cell association and interference coordination in heterogeneous LTE-A cellular networks,” *IEEE Journal on Selected Areas in Communications (JSAC)*, 2010.

[2] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, “LTE-Advanced: Heterogeneous networks,” in *Proc. European Wireless Conference*, 2010.

[3] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, “A survey on 3GPP heterogeneous networks,” *IEEE Wireless Communications*, 2011.

[4] T. Bonald and A. Proutiere, “Wireless downlink data channels: User performance and cell dimensioning,” in *Proc. Mobile Computing and Networking (MobiCom)*, 2003.

[5] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, “An overview of load balancing in HetNets: Old myths and open problems,” *IEEE Wireless Communications*, 2014.

[6] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, “Providing Quality of Service over a shared wireless link,” *IEEE Communications Magazine*, 2001.

[7] A. Jalali, R. Padovani, and R. Pankaj, “Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system,” in *Proc. Vehicular Technology Conference*, 2000.

[8] P. Hande, S. Patil, and H. Myung, “Distributed load-balancing in a multi-carrier wireless system,” in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2009.

[9] 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.

[10] H. Boostanimehr and V. Bhargava, “Unified and distributed QoS-driven cell association algorithms in heterogeneous networks,” *IEEE Transactions on Wireless Communications*, 2015.

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

[12] 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 Communications (JSAC)*, 2011.

[13] T. Han and N. Ansari, “Smart grid enabled mobile networks: Jointly optimizing BS operation and power distribution,” in *Proc. IEEE International Conference on Communications*, 2014.

[14] J. Bartelt, A. Fehske, H. Klessig, G. Fettweis, and J. Voigt, “Joint bandwidth allocation and small cell switching in heterogeneous networks,” in *Proc. IEEE Vehicular Technology Conference*, 2013.

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

[16] 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.

[17] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, “Reducing the energy consumption of small cell networks subject to QoE constraints,” in *Proc. IEEE Globecom*, 2014.

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

[19] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, “The evolution of cellular backhaul technologies: Current issues and future trends,” *in IEEE Communications Surveys and Tutorials*, 2011.

[20] Y. Wang and K. Pedersen, “Performance analysis of enhanced inter-cell interference coordination in LTE-Advanced heterogeneous networks,” in *Vehicular Technology Conference (VTC Spring)*, 2012.

[21] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, “Algorithms for enhanced inter-cell interference coordination (eICIC) in lte hetnets,” *IEEE/ACM Transasctions on Networking*, 2014.

[22] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, “Coordinated multipoint transmission and reception in LTE-advanced systems,” *IEEE Communications Magazine*, 2012.

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

[24] 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.

[25] O. Somekh, O. Simeone, A. Sanderovich, B. Zaidel, and S. Shamai, “On the impact of limited-capacity backhaul and inter-users links in cooperative multicell networks,” in *Proc. Conference Information Sciences and System (CISS)*, 2008.

[26] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, “Cloud technologies for flexible 5G radio access networks,” *IEEE Communications Magazine*, 2014.

[27] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, “Joint uplink and downlink cell selection in cognitive small cell heterogeneous networks,” in *Proc. IEEE Globecom*, 2014.

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

[29] G. T. . v.12.0.0 Rel.12, *Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects*. Academic press, 2013.

[30] H. Kim, H. Y. Kim, Y. Cho, and S.-H. Lee, “Spectrum breathing and cell load balancing for self organizing wireless networks,” in *Proc. IEEE Communications Workshops*, 2013.

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

[32] T. Bonald, S. Borst, N. Hegde, M. Jonckheere, and A. Proutiere, “Flow-level performance and capacity of wireless networks with user mobility,” 2009.

[33] 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.

[34] “http://cbnl.com/solutions-mobile-backhaul.”

[35] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, “Optimal downlink and uplink user association in backhaul-limited hetnets,” in *(to appear) IEEE Infocom*, 2016.

[36] 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.

[37] G. 36.300, “Evolved universal terrestrial radio access (E-UTRA); further enhancements to LTE time division duplex (TDD) for downlink-uplink (DL-UL) interference management and traffic adaptation,” 2012.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

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

[40] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, “Energy-efficient user association in cognitive heterogeneous networks,” *IEEE Communications Magazine*, 2014.

[41] *3GPP, Technical Report LTE; Evolved Universal Terrestrial Radio Access (E-UTRA)*, TR 136 931, 2011.

[42] Alcatel-Lucet, “Mobile bakhaul architecture for hetnet, https://www.alcatel-lucent.com/solutions/mobile-backhaul,” 2015.

[43] G. 36.828, “Evolved universal terrestrial radio access (E-UTRA) and radio access network (E-UTRAN); overall descriptiom,” 2012.

[44] T. Bu, L. Li, and R. Ramjee, “Generalized proportional fair scheduling in third generation wireless data networks,” in *Proc. IEEE Infocom*, 2006.

[45] S. Das, H. Viswanathan, and G. Rittenhouse, “Dynamic load balancing through coordinated scheduling in packet data systems,” in *Proc. IEEE Infocom*, 2003.

[46] I. S. 802.16m, “IEEE p802.16m-2007 draft standards for local and metropolitan area networks part 16: Air interface for fixed broadcast wireless access systems„” 2007.

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